VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra aplikované matematiky

Normalita dat

Normality of Data

2016

Lucie Matějková

VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra aplikované matematiky

Zadání bakalářské práce

Student: **Lucie Matějková**

Studijní program: B2647 Informační a komunikační technologie

Studijní obor: 1103R031 Výpočetní matematika

Téma: Normalita dat
Normality of Data

Jazyk vypracování: čeština

Zásady pro vypracování:

Metody klasické analýzy dat jsou založeny na předpokladu normality. Před jejich aplikací se musíme přesvědčit, zda pozorování představují realizaci náhodného výběru pocházejícího z normálního rozdělení, a to buď vizuálně nebo použitím testu normality. Cílem práce je popis způsobů ověřování normality dat, srovnání testů normality a generování volně šiřitelného výpočetního appletu pro ověření normality. Formální náležitosti a rozsah bakalářské práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Postup práce:
1. Grafické metody pro ověření normality.
2. Testy normality.
3. Srovnání testů normality (srovnání chyby I. a II. druhu na základě simulací).
4. Transformace dat vedoucí k přiblížení k normalitě.
5. Generování volně šiřitelného výpočetního appletu pro ověřování normality.

Seznam doporučené odborné literatury:

1. ÖZTUNA D., ELHAN A. H., TÜCCAR E. (2006), Investigation of Four Different Normality Tests in Terms of Type 1 Error Rate and Power under Different Distributions, Turkish Journal of Medical Sciences.
2. YAP B. W., SIM C. H. (2011), Comparisons of various types of normality tests, Journal of Statistical Computation and Simulation, Volume 81, Issue 12, pg. 2141 – 2155.
3. LITSCHMANNOVÁ, M. (2011), Úvod do statistiky, Učební texty v elektronické podobě, dostupné z Word Wide Web: http://mi21.vsb.cz/modul/uvod-do-statistiky

Formální náležitosti a rozsah bakalářské práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí bakalářské práce: **Ing. Martina Litschmannová, Ph.D.**

Datum zadání: 01.09.2015

Datum odevzdání: 29.04.2016

doc. RNDr. Jiří Bouchala, Ph.D.
vedoucí katedry

prof. RNDr. Václav Snášel, CSc.
děkan fakulty

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně. Uvedla jsem všechny literární prameny a publikace, ze kterých jsem čerpala.

V Ostravě 16. dubna 2016

.....

Ráda bych na tomto místě poděkovala Ing. Martině Litschmannově, Ph.D., za odborné vedení mé bakalářské práce, cenné rady a připomínky, stejně tak za materiály, které mi poskytla, ale především za čas, který mi věnovala.

Abstrakt

Cílem mé práce, která je zaměřena na normalitu dat, je jak popsání grafických metod, tak popsání statistických testů zabývajících se normalitou. Nejprve jsou uváděny základní informace o normálním rozdělení pravděpodobnosti. Dále popisujeme grafické metody, které se používají při testování normality, jsou to metody Q-Q graf, P-P graf a N-P graf. Poté pokračujeme statistickými testy, kde uvádíme Chí-kvadrát test dobré shody, Lillieforsův Kolmogorovův-Smirnovův test, testy založené na šikmosti a špičatosti, Shapirův-Wilkův test a nakonec Andersonův-Darlingův test. Nejdůležitější částí je simulační studie prováděna pro srovnání těchto pěti testů podle chyb I. a II. druhu. Výběry z různých typů rozdělení pravděpodobnosti byly generovány programem RStudio. Ve stejném programu byly prováděny i testy a simulační studie pro srovnání testů.

Klíčová slova: normalita dat, normální rozdělení, testy normality, Chí-kvadrát test, Lillieforsův Kolmogorovův-Smirnovův test, šikmost, špičatost, Shapirův-Wilkův test, Andersonův-Darlingův test

Abstract

The object od my work is description od graphic methods, also description of statistical tests which deal with normality. At first I desribe basic information about normal distribution of probability. Then I describe graphic mothods, which are used during tests of normality. The methods are Q-Q plot, P-P plot and N-P plot. After that I continue with statistical tests, where I describe Chi-squared test, Lilliefors Kolmogorov-Smirnov test, tests based on skewness and kurtosis, Shapiro-Wilk test and Anderson-Darling test. The most important part is simulation study made to compare these five tests according to errors I. and II. type. Selections from different types of probability distribution were generated by software RStudio. Tests and simulation studies for comparison of tests were made in the same software.

Keywords: normality of data, normal distribution, tests of normality, Chi-squared test, Lilliefors Kolmogorov-Smirnov test, skewness, kurtosis, Shapiro-Wilk test, Anderson-Darling test

Seznam použitých zkratek a symbolů

AD	–	Testová statistika Andersonovova-Darlingova testu
$AD_{0,95}$	–	Kritická hodnota Andersonovova-Darlingova testu
D_n	–	Testová statistika Lillieforsova Kolmogorovova-Smirnovova testu
$D_n(\alpha)$	–	Kritická hodnota Lillieforsova Kolmogorovova-Smirnovova testu
$D(X)$ nebo σ^2	–	Rozptyl
$E(X)$ nebo μ	–	Střední hodnota
$F(x)$	–	Distribuční funkce
U_3	–	Testová statistika testu šikmosti
U_4	–	Testová statistika testu špičatosti
W	–	Testová statistika Shapirovova-Wilkova testu
X, Y, Z	–	Náhodná veličina
$f(x)$	–	Hustota pravděpodobnosti
α	–	Hladina významnosti, pravděpodobnost chyby I. druhu
β	–	Pravděpodobnost chyby II. druhu
σ	–	Směrodatná odchylka
$\phi(z)$	–	Charakteristická funkce normovaného normálního rozdělení
$\varphi(z)$	–	Hustota pravděpodobnosti normovaného normálního rozdělení
χ^2	–	Testová statistika Chí-kvadrát testu dobré shody
$\psi(x)$	–	Charakteristická funkce
$\Phi(z)$	–	Distribuční funkce normovaného normálního rozdělení

Obsah

1 Úvod **1**

2 Základní pojmy a teorie pravděpodobnosti **2**
2.1 Základní pojmy 2
2.2 Číselné charakteristiky 5
2.3 Testování hypotéz 7

3 Normální rozdělení **10**
3.1 Normované normální rozdělení 12

4 Grafické metody pro ověření normality dat **17**
4.1 Srovnání empirické a teoretické hustoty 17
4.2 Kvantilově-kvantilový graf (Q-Q graf) 18
4.3 Pravděpodobnostní graf (P-P graf) 20
4.4 Normální pravděpodobnostní graf (N-P graf) 22

5 Testy normality **25**
5.1 Chí-kvadrát test dobré shody 25
5.2 Lillieforsův (Kolmogorovův-Smirnovův) test 26
5.3 Testy založené na šikmosti a špičatosti 27
5.4 Shapirův-Wilkův test 28
5.5 Andersonův-Darlingův test 29

6 Srovnání testů normality **30**
6.1 Srovnání podle chyby I. druhu 30
6.2 Srovnání podle chyby II. druhu 33

7 Transformace dat vedoucí k přiblížení k normalitě **40**
7.1 Boxova-Coxova transformace 40

8 Volně šířitelný software pro ověřování normality **43**

9 Závěr **44**

10 Reference **45**

Přílohy **45**

A Funkce pro test založený na šikmosti a špičatosti **46**

B Funkce pro srovnání testů na základě chyby I. druhu **47**

C Funkce pro srovnání testů na základě chyby II. druhu **48**

D Distribuční funkce normovaného normálního rozdělení pro $z > 0$ **49**

Seznam tabulek

1 Přehled možných výsledků testování hypotéz a jejich pravděpodobností . 9
2 Výběrové charakteristiky pro odhad pravděpodobnosti chyby I. druhu . . 32
3 Výsledky Shapirova-Wilkova testu před a po transformaci 42
4 Výsledky Shapirova-Wilkova testu před a po transformaci 42

Seznam obrázků

2.1 Příklad grafu distribuční funkce spojité náhodné veličiny 4
2.2 Šikmost a špičatost u vybraných typů rozdělení 7
3.1 Gaussova křivka 11
3.2 Hustota pravděpodobnosti náhodné veličiny s rozdělením $N(\mu = 20; \sigma)$. 11
3.3 Hustota pravděpodobnosti náhodné veličiny s rozdělením $N(\mu; \sigma = 5)$ . . 12
3.4 Hustota pravděpodobnosti normovaného normálního rozdělení 13
4.1 Porovnání empirické a teoretické hustoty pravděpodobnosti 18
4.2 Grafické zobrazení zešikmeného rozdělení pomocí Q-Q grafů 19
4.3 Grafické zobrazení Q-Q grafů pro náhodné výběry z normálního a rovnoměrného rozdělení 20
4.4 Grafické zobrazení zešikmeného rozdělení pomocí P-P grafů 21
4.5 P-P grafy pro náhodné výběry z normálního a rovnoměrného rozdělení . 22
4.6 Grafické zobrazení zešikmeného rozdělení pomocí N-P grafů 23
4.7 Grafické zobrazení N-P grafů z náhodných výběrů normálního a rovnoměrného rozdělení 24
6.1 Odhad pravděpodobnosti chyby I. druhu při rozhodování nulové hypotézy o normalitě dat 31
6.2 Vícenásobný krabicový graf pro odhady pravděpodobností chyby I. druhu 33
6.3 Odhad pravděpodobnosti chyby II. druhu při rozhodování hypotézy o normalitě dat výběru z rovnoměrného rozdělení 34
6.4 Odhad pravděpodobnosti chyby II. druhu při rozhodování hypotézy o normalitě dat výběru z exponenciálního rozdělení 35
6.5 Odhad pravděpodobnosti chyby II. druhu při rozhodování hypotézy o normalitě dat výběru z logaritmicko-normálního rozdělení 36
6.6 Odhad pravděpodobnosti chyby II. druhu při rozhodování hypotézy o normalitě dat výběru ze Studentova rozdělení s jedním stupněm volnosti 37
6.7 Odhad pravděpodobnosti chyby II. druhu při rozhodování hypotézy o normalitě dat výběru ze Studentova rozdělení se třemi stupni volnosti . . 38
6.8 Odhad pravděpodobnosti chyby II. druhu při rozhodování hypotézy o normalitě dat výběru ze Studentova rozdělení s deseti stupni volnosti . . 39
7.1 Boxova-Coxova transformace dat kladně zešikmeného rozdělení na normální rozdělení 41

Seznam výpisů zdrojového kódu

1 Funkce pro testování šikmosti a špičatosti 46
2 Funkce pro testování testů na základě chyby I. druhu 47
3 Funkce pro testování testů na základě chyby II. druhu 48

1 Úvod

V této bakalářské práci se budeme zabývat převážně normálním rozdělením pravděpodobnosti a testováním normality datových souborů, jak z pohledu grafických metod, tak pomocí statistických testů pro ověření normality dat.

V první části mé práce se věnuji zavedení základních pojmů z oblasti pravděpodobnosti, uvádím zde pojmy spjaty s číselnými charakteristikami a testováním hypotéz. Další důležitou částí je popis normálního rozdělení pravděpodobnosti a normovaného normálního rozdělení.

Nádledující dvě kapitoly jsou věnovány ověřování normality. Jedna z nich je zaměřena na grafické ověření normality datových výběrů, které nám mohou před samotným testováním prozradit, co můžeme od jednotlivých testů očekávat. Pomocí několika typů grafů, jako je Q-Q graf, P-P graf a N-P graf, jsme demonstrovali různá rozdělení pravděpodobnosti a posuzovali normalitu. V další kapitole testujeme datové výběry pomocí různých statistických testů pro ověření normality dat, a zjišťujeme, jak testy fungují. V naší práci studujeme statistické testy pro ověření normality dat, a to jsou Chí-kvadrát test dobré shody, test Lillieforsův Kolmogorovův-Smirnovův, dále testy založené na šikmosti, testy založené na špičatosti a testy založené na šikmosti i špičatosti zároveň, předposledním testem je Shapirův-Wilkův test a nakonec Andersonův-Darlingův test normality.

Nejdůležitější částí bakalářské práce jsou simulační studie, ve kterých provádíme srovnání výše zmíněných testů normality na základě chyb I. a II. druhu.

Pro vygenerování různých typů rozdělení pravděpodobnosti, sestrojení grafů a použití statistických testů jsme využili softwaru RStudio.

2 Základní pojmy a teorie pravděpodobnosti

Abychom mohli charakterizovat normální rozdělení, potřebujeme si zavést několik základních pojmů z oblasti pravděpodobnosti.

2.1 Základní pojmy

Nejprve si řekneme co je to náhodný pokus, základní prostor, elementární jev, náhodný jev a náhodná veličina.

Děj, jehož výsledek není předem jednoznačně určen podmínkami, za kterých probíhá je označován jako *náhodný pokus*. Pro matematický popis náhodného pokusu určujeme *základní prostor* Ω, který je množinou všech dále nedělitelných výsledků daného pokusu.

Prvky množiny Ω, popřípadě jednoprvkové podmnožiny Ω nazýváme *elementárními jevy* a označujeme je $\{\omega\}$. *Náhodný jev* představuje událost, která za určitých podmínek buď nastane nebo nenastane. Náhodným jevem považujeme každou podmnožinu základního prostoru Ω. Pro náhodné jevy platí stejné rovnosti a algebraické zákony jako pro výroky. Pro označení se používají velká písmena abecedy (např.: A, B, ...).

Za *pravděpodobnost* považujeme míru předpokládatelnosti výskytu náhodného jevu. Nabývá hodnot z intervalu $< 0, 1 >$. Čím vyšší pravděpodobnost je, tím je vyšší i šance, že náhodný jev nastane. Jev, který nastane nutně při každém provedení náhodného pokusu je *jev jistý* a má pravděpodobnost 1. Opak jistého jevu je *jev nemožný* s pravděpodobností 0, jedná se o jev, který v pokusu nemůže nikdy nastat.

Definice 2.1.1 *Náhodná veličina X je reálná funkce $X : \Omega \to \mathbb{R}$ taková, že pro každé reálné x je množina*

$$\{\omega \in \Omega | X(\omega) < x\}$$

náhodným jevem.

Za hodnotu *náhodné veličiny* X budeme považovat výsledek náhodného pokusu vyjádřený reálným číslem. Náhodné veličiny jsou libovolné veličiny, které můžeme opakovaně měřit u odlišných objektů.

Náhodná veličina může být například doba do poruchy nějakého přístroje, počet vadných produktů mezi n produkty, změřená teplota na daném místě ve stejnou dobu, ale jiný den, roční mzda očanů města, atd.

Definice 2.1.2 *Nechť X je náhodná veličina. Reálnou funkci $F(x)$ definovanou pro všechna reálná x vztahem*

$$F(x) = P(X < x)$$

nazýváme distribuční funkcí náhodné veličiny X.

Z definice 2.1.2 vyplývá řada vlastností distribuční funkce:

1. $0 \leq F(x) \leq 1$, tzn., že funkce $F(x)$ nabývá hodnot z intervalu $(0, 1\rangle$,
2. $\forall x_1, x_2, x_1 < x_2 : F(x_1) \leq F(x_2)$, jedná se o funkci neklesající,
3. $\forall a \in \mathbb{R} : \lim_{x \to a^+} F(x) = F(a)$, funkce $F(x)$ je zleva spojitá,
4. distribuční funkce má nejvýše spočetně mnoho bodů nespojitosti,
5. $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$. [9]

Distribuční funkce slouží pro popis tzv. *dikretní i spojité náhodné veličiny.*

Definice 2.1.3 *Řekněme, že náhodná veličina X se nazývá diskrétní právě tehdy, když nabývá nejvýše spočetně mnoha hodnot $\{x_1, x_2, ...\}$ tak, že*

$$P(X = x_i) \geq 0; \qquad \sum_{i=1}^{\infty} P(X = x_i) = 1.$$

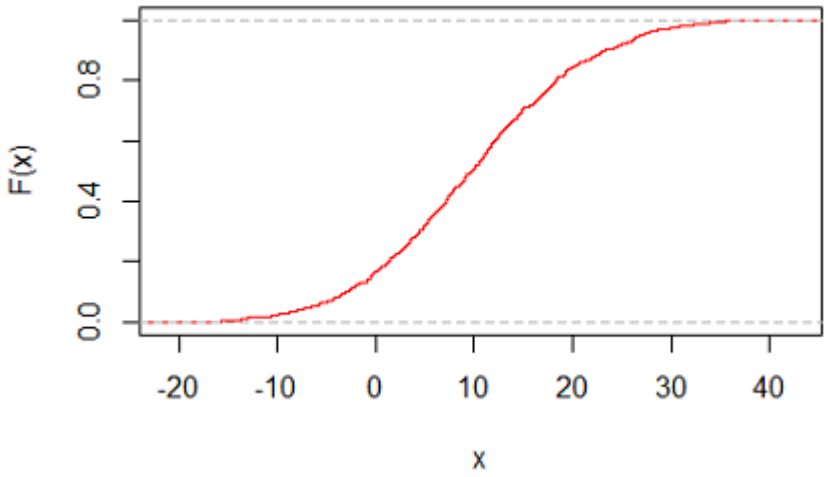
Nejčastěji se jedná o veličiny celočíselné. Příkladem diskrétní náhodné veličiny je počet členů domácnosti, počet šestek při deseti hodech kostkou, počet dopravních nehod v Ostravě za jeden den, atd.

Definice 2.1.4 *Řekneme, že náhodná veličina X se nazývá spojitá právě tehdy, když může nabývat všech hodnot spojité distribuční funkce.*

Jako příklad spojité náhodné veličiny lze uvést životnost výrobku, naměřenou hodnotu napětí, náhodně vybrané reálné číslo, atd.

Dále si zavedeme pojem *rozdělení pravděpodobnosti náhodné veličiny*. Rozdělení náhodné veličiny X nám charakterizuje, jakých hodnot může náhodná veličina X nabývat a s jakými pravděpodobnostmi. Rozdělení pravděpodovnosti náhodné veličiny se dělí na diskrétní a spojité rozdělení. V našem případě se budeme věnovat spojitému rozdělení náhodné veličiny.

Z Definice 2.1.4 plyne, že v případě spojité náhodné veličiny nemá smysl přiřazovat hodnotu pravděpodobnosti jednotlivým realizacím spojité náhodné veličiny. Přičemž na libovolném intervalu pravděpodobnost výskytu vymezit můžeme. To značí, že pro popis můžeme použít distribuční funkci spojité náhodné veličiny, Obrázek 2.1.



Obrázek 2.1: Příklad grafu distribuční funkce spojité náhodné veličiny

Kromě distribuční funkce se používá k popisu spojité náhodné veličiny tzv. *hustota pravděpodobnosti* $f(x)$.

Definice 2.1.5 *Hustota pravděpodobnosti $f(x)$ spojité náhodné veličiny je reálná nezáporná funkce taková, že distribuční funkci $F(x)$ lze vyjádřit ve tvaru $F(x) = \int_{-\infty}^{x} f(t)dt, -\infty < x < \infty$.*

V Definice 2.1.5 je distribuční funkce spojitá pro všechna x ve všech bodech, kde má derivaci. Základní vlastnosti hustoty pravděpodobnosti náhodné veličiny x jsou:

1. $f(x) \geq 0, pro -\infty < x < \infty$, hustota pravděpodovnosti je nezáporná funkce,
2. $\int_{-\infty}^{\infty} f(x)dx = 1$, tzn., že plocha pod křivkou hustoty pravděpodovnosti je rovna 1,
3. $\lim_{x\to-\infty} f(x) = 0$, $\lim_{x\to\infty} f(x) = 0$. [9]

Z Definice 2.1.5 hustoty pravděpodobnosti lze rovněž snadno odvodit vztahy mezi pravděpodobností a hustotou pravděpodobnosti.

1. $P(X < a) = F(a) = \int_{-\infty}^{a} f(x)dx$, pro všechna $a \in \mathbb{R}$
2. $P(X \geq a) = 1 - F(a) = \int_{a}^{\infty} f(x)dx$, pro všechna $a \in \mathbb{R}$
3. $P(a \leq X < b) = F(b) - F(a) = \int_{a}^{b} f(x)dx$, pro všechna $a < b$; $a, b \in \mathbb{R}$. [9]

Budeme také potřebovat charakteristickou funkci spojitého rozdělení náhodné veličiny, protože později použijeme tuto fukci při důkazu Linderbergovy-Lévyho věty.

Definice 2.1.6 *Nechť X je náhodná veličina. Pak funkce $\psi : \mathbb{R} \to \mathbb{C}$ daná vztahem $\psi_X(x) = E(e^{ixX})$, $x \in \mathbb{R}$, se nazývá charakteristickou funkcí náhodné veličiny X.*

2.2 Číselné charakteristiky

Rozdělení pravděpodobnosti náhodné veličiny X je popsáno pomocí její distribuční funkce $F(x)$, popřípadě pomocí hustoty pravděpodobnosti $f(x)$ nebo charakteristickou funkcí $\psi(x)$. Ve většině případů je výhodné shrnout celkovou informaci o této náhodné veličině do několika čísel. Tato čísla nám charakterizují některé vlastnosti náhodné veličiny a říkáme jim číselné charakteristiky náhodné veličiny X. Představíme si pouze ty pro nás důležité.

2.2.1 Střední hodnota a rozpyl

Máme-li náhodnou veličinu X se spojitým rozdělením, její *střední hodnota* je číslo

$$E(X) = \mu = \int_{-\infty}^{\infty} x \cdot f(x)dx. \tag{2.1}$$

Střední hodnota má smysl jen, pokud integrál (2.1) existuje. Je nejznámější mírou polohy ve statistice.

Rozptyl náhodné veličiny X je číslo

$$D(X) = \sigma^2 = \int_{-\infty}^{\infty} (x - E(X))^2 \cdot f(x)dx. \tag{2.2}$$

Stejně jako u střední hodnoty, rozptyl existuje, pokud existuje integrál (2.2). Jedná se o druhý centrální moment náhodné veličiny, který vyjadřuje variabilitu rozdělení souboru náhodných hodnot kolem její střední hodnoty.

Směrodatnou odchylku určuje číslo

$$\sigma = \sqrt{D(X)}. \tag{2.3}$$

Podobně jako rozptyl, určuje směrodatná odchylka jak moc jsou hodnoty rozptýleny či odchýleny od průměru hodnot. Jedná se o odmocninu z rozptylu.

2.2.2 Míry šikmosti a špičatosti

Charakteristiky, které se používají méně často, ale za to obvykle společně slouží k vystižení dalších vlastností hodnot souboru. Pomocí šikmosti a špičatosti hodnotíme, jak se rozdělení dat podobá tzv. *normálnímu rozdělení*.

Nejprve si musíme zavést pojem *centrální moment k-tého řádu*, který označujeme jako ${\mu_k}'$, pro $k = 1, 2, 3, \cdots$. Pro naší práci budeme potřebovat znát pouze centrální moment spojité náhodné veličiny, který je definován

$${\mu_k}' = \int_{-\infty}^{\infty} (x - E(X))^k \cdot f(x)dx. \tag{2.4}$$

Šikmost je mírou symetrie daného rozdělení a měří zešikmenost, resp. nesymetrii dat. Šikmostí tedy zkoumáme, jestli jsou hodnoty rozloženy okolo průměru symetricky. Definujeme ji podílem třetího centrálního momentu a třetí mocniny směrodatné odchylky.

$$a_3 = \frac{\mu_3}{\sigma^3} \tag{2.5}$$

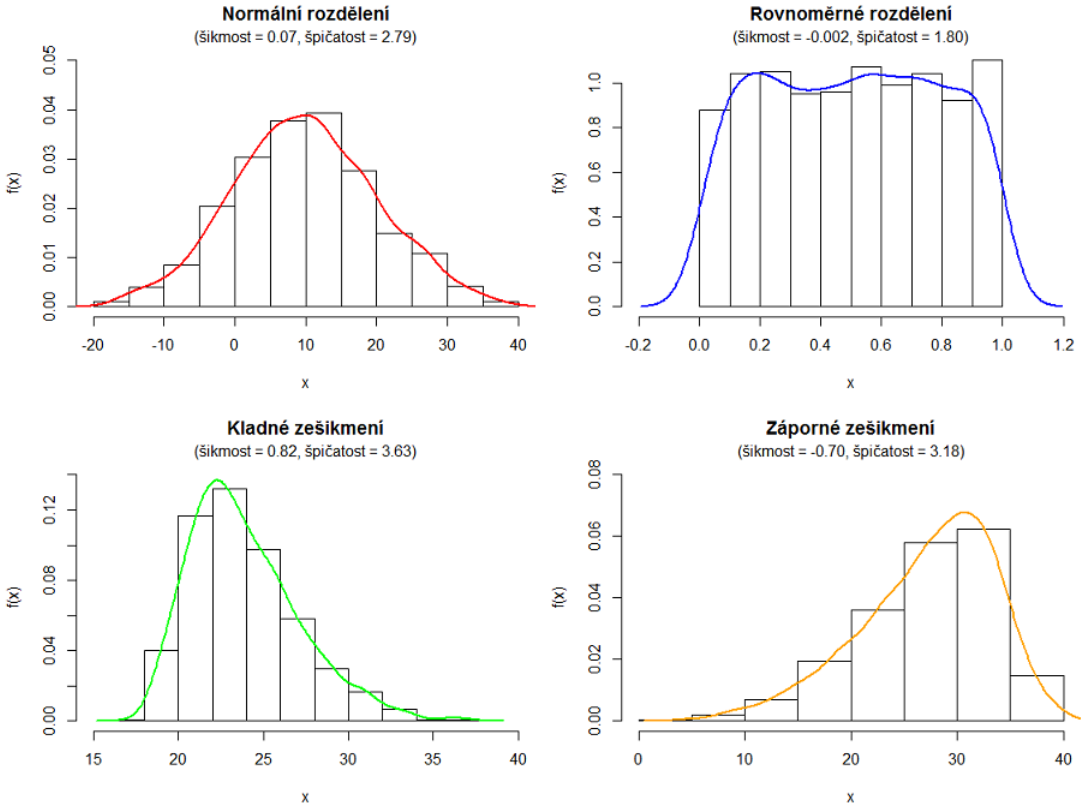
Pro symetrické rozdělení platí, že $a_3 = 0$, pro rozdělení pozitivně zešikmené $a_3 > 0$ a pro rozdělení, které je negativně zešikmené $a_3 < 0$.

Špičatost je mírou koncentrace hodnot náhodné veličiny kolem střední hodnoty a měří odchylku špičatosti zkoumaného rozdělení od normálního rozdělení. Definujeme ji podílem čtvrtého centrálního momentu a čtvrté mocniny směrodatné odchylky.

$$a_4 = \frac{\mu_4}{\sigma^4} \tag{2.6}$$

Pro špičatost normálního rozdělení platí, že $a_4 = 3$, pro větší špičatost než u normálního rozdělení je $a_4 > 3$ a naopak pro menší špičatost než u normálního rozdělení je $a_4 < 3$.

Jak se mění šikmost a špičatost pro vybrané typy rozdělení pravděpodobnosti nám zobrazuje Obrázek 2.2.



Obrázek 2.2: Šikmost a špičatost u vybraných typů rozdělení

2.3 Testování hypotéz

Testování hypotéz je proces pro ověřování správnosti vyslovené hypotézy pomocí výsledků získaných z výběrového zkoumání statistické hypotézy. *Statistická hypotéza* je tvrzením pojednávajícím o rozdělení pozorované náhodné veličiny X. Dělí se na parametrickou a neparametrickou hypotézu. Statistická *parametrická hypotéza* vypovídá i o parametrech rozdělení náhodné veličiny jako je např. střední hodnota, rozptyl, pravděpodobnost a jiné. Jakožto *neparametrická hypotéza* se týká jiných vlastností náhodné veliřiny.

2.3.1 Nulová a alternativní hypotéza

Testování hypotéz je tzv. rozhodovací proces, v němž stojí proti sobě dvě tvrzení (nulová a alternativní hypotéza).

1. *Nulová hypotéza* H_0 je tvrzení, které bývá obvykle vyjádřeno rovností mezi testovaným parametrem a jeho očekávanou hodnotou. $H_0 : \theta = \theta_0$

2. *Alternativní hypotéza* H_1 popírá tvrzení dané nulovou hypotézou. V připadě jednovýběrových testů ji můžeme zapsat jedním z několika způsobů.

   - $H_1 : \theta = \theta_1$, tzv. jednoduchá alternativní hypotéza, použijeme ji v případě, kdy rozhodujeme mezi dvěmi hodnotami θ_0 a θ_1.
   - $H_1 : \theta \neq \theta_0$, alternativní hypotéza složená, popírá platnost hypotézy nulové bez jakékoli bližší specifikace. Takhle formulovaná hypotéza se nazývá oboustranná.
   - $H_1 : \theta > \theta_0$, patří mezi složené alternativní hypotézy. Alternativní hypotéza formulovaná jako jednostranná, popírá nulovou hypotézu a zároveň tvrdí, že hodnota testovaného parametru je větší než hodnota uvedená v nulové hypotéze.
   - $H_1 : \theta < \theta_0$, stejně jako předchozí alternativní hypotéza, patří mezi složené hypotézy a je jednostranná. Tato hypotéza popírá nulovou a zároveň tvrdí, že hodnota testového parametru je menší než uvedená hodnota v nulové hypotéze.

Testem statistické hypotézy je jakýsi postup, při němž na základě výběrového souboru provádíme rozhodnutí, která z předpokládaných hypotéz uspěje. Hypotézy proto musíme formulovat tak, aby při rozhodnutí uspěla právě jedna. Při testu statistické hypotézy, který se dá provádět opakovaně, je pochopitelné, že můžeme dojít ke dvěma rozhodnutím.

1. Nezamítáme hypotézu H_0.

2. Zamítáme H_0 ve prospěch hypotézy H_1.

Ke kterému z rozhodnutí se přiklonit nám určuje obor hodnot testovaného parametru θ, který se dělí na dvě disjunktní množiny nazývané *obor přijetí* V hypotézy H_0 a *kritický obor* W - obor zamítnutí hypotézy H_0. Hranice mezi těmito obory se nazývá *kritická hodnota testu* t_{krit}.

K provedení konkrétního testu statistické hypotézy musíme mít k dispozici testovou statistiku (někdy také testové kritérium), kterou je výběrová charakteristika $T(X)$. Tato testová statistika má vztah k H_0 a její rozdělení pravděpodobnosti za předpokladu platnosti nulové hypotézy známe.

Kritický obor W jde často popsat pomocí kritického oboru W^* testové statistiky $T(X)$. Pokud tedy padne pozorovaná hodnota testové statistiky do kritického oboru W^*, zamítáme nulovou hypotézu H_0, pokud padne do oboru přijetí V, nezamítáme hypotézu H_0.

2.3.2 Chyba I. a II. druhu

		Výsledek testu	
		Nezamítáme H_0	Zamítáme H_0
Skutečnost	Platí H_0	Správné rozhodnutí $1-\alpha$ (spolehlivost testu)	**Chyba I. druhu** α (hladina významnosti)
	Platí H_1	**Chyba II. druhu** β	Správné rozhodnutí $1-\beta$ (síla testu)

Tabulka 1: Přehled možných výsledků testování hypotéz a jejich pravděpodobností Tabulka převzatá z [10].

Při rozhodování můžeme dojít k jednomu ze závěrů, který je uveden v Tabulka 1.

Jestliže platí nulová hypotéza a my jsme ji nezamítli, rozhodli jsme se správně. Pravděpodobnost našeho rozhodnutí označujeme $1-\alpha$ a nazýváme ji *spolehlivost testu*. Pokud je nulová hypotéza H_0 ve skutečnosti platná, ale my ji i přesto zamítáme, dopouštíme se *chyby I. druhu*. Pravděpodobnost, že k tomuhle pochybení dojde, nazýváme *hladinou významnosti* α. Platí-li hypotéza alternativní a my jsme rozhodli o zamítnutí H_0, usoudili jsme opět správně. Takovéto rozhodnutí má pravděpodobnost $1-\beta$, kterou označujeme jako *sílu testu*. Poslední možností při rozhodování je, když zamítneme alternativu přestože je vlastně platná, tím se dopouštíme *chyby II. druhu*. Pravděpodobnosti chyby je označována β.

Ve statistice volíme jako rozhodující vstupní parametr testu hladinu významnosti α, která značí pravděpodobnost chyby I. druhu. V technických oblastech obvykle volíme hladinu významnosti $\alpha = 0,05$, ve speciálních případech nároky na pravděpodobnost chyby I. druhu ještě zvyšujeme (volíme $\alpha = 0,01$, např. v biostatistice).

Chybu II. druhu β snižujeme volbou vhodného testu (pokud máme možnost výběru), popřípadě zvětšením rozsahu výběrového souboru, což je jediný způsob jak snížit pravděpodobnost chyby II. druhu β, aniž bychom tím zvýšili pravděpodobnost chyby I. druhu α.

3 Normální rozdělení

S normálním rozdělením se setkáváme nejčastěji a je jedním z nejdůležitějších spojitých rozdělení pravděpodobnosti, které popisuje celou řadu veličin. Normální rozdělení je také známé jako Gaussovo rozdělení. Jedná se o jedno z nejpoužívanějších a nejdůležitějších pravděpodobnostních rozdělení spojité náhodné veličiny. Normální rozdělení modeluje velké množství chování náhodných veličin, které se vyskytují ve společnosti nebo v přírodě.

Takové náhodné veličiny, řídící se normálním rozdělením, jsou například výška lidí v populaci nebo IQ populace, vitální kapacita plic, odchylka rozměru produktu od požadované hodnoty, velikost chyby měření.

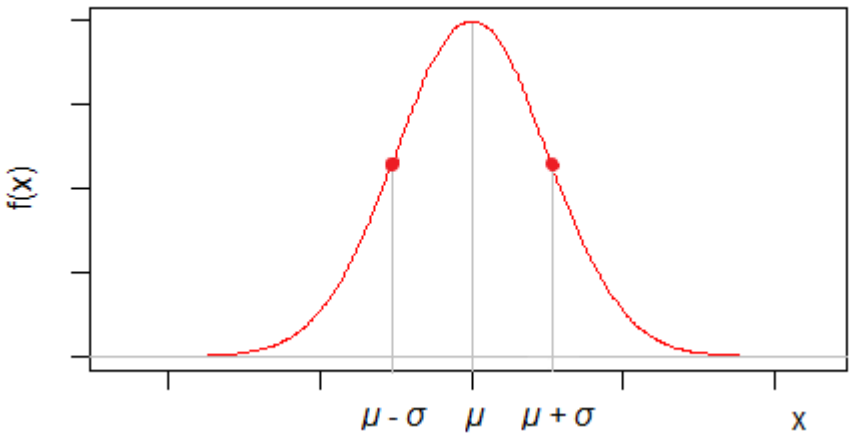
Původně bylo normální rozdělení odvozeno pro analýzu chyb měření, způsobených velkým počtem vzájemně nezávislých a neznámých příčin, ale postupně se ukázalo, že za určitých podmínek toto rozdělení dobře aproximuje spoustu jiných pravděpodobnostních rozdělení, jak spojitých, tak i diskrétních.

Fakt, že náhodná veličina X má normální rozdělení, zapisujeme $X \sim N(\mu; \sigma^2)$, kde μ je střední hodnota, která charakterizuje polohu rozdělení a σ^2 je rozptyl, určující rozptýlení hodnot náhodné veličiny kolem střední hodnoty.

Každou spojitou náhodnou veličinu lze popsat například hustotou pravděpodobnosti, distribuční funkcí nebo charakteristickou funkcí. Tyto funkce jsou ekvivalentní v tom smyslu, že jedna se dá spočítat z druhé.

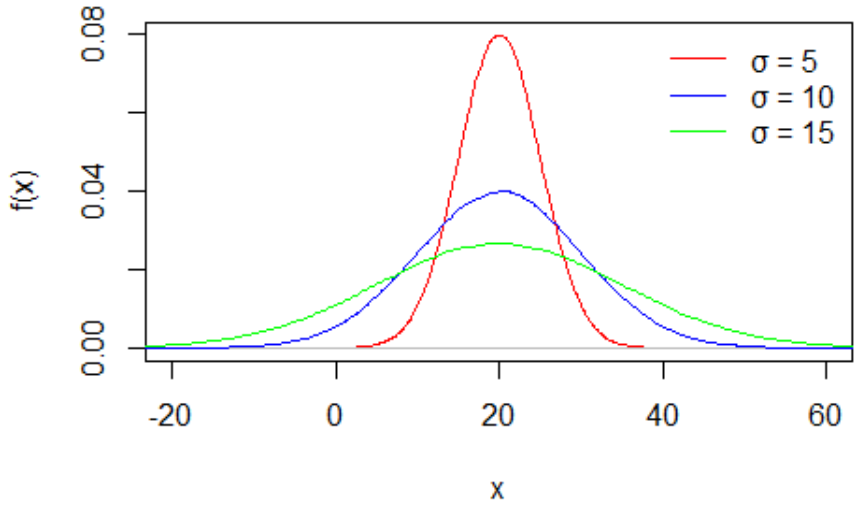
Hustota pravděpodobnosti rozdělení normální veličiny X je dána vztahem

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \text{pro} -\infty < x < \infty. \tag{3.1}$$



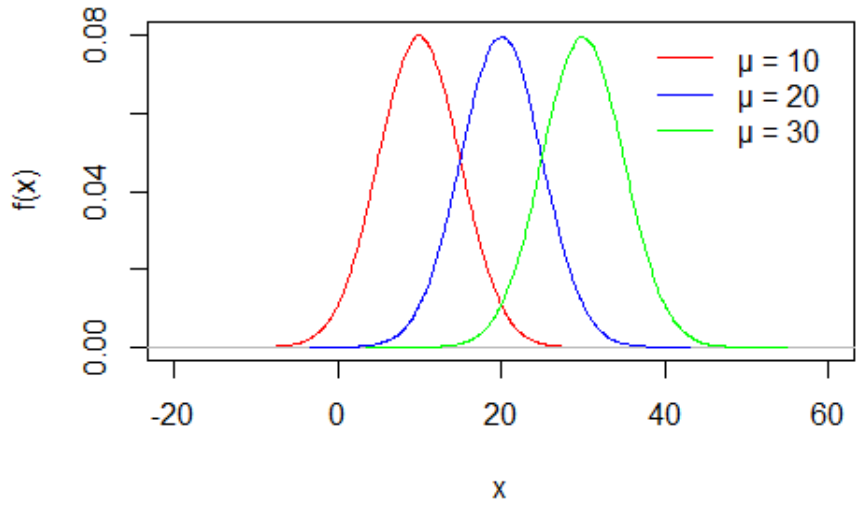
Obrázek 3.1: Gaussova křivka

Grafickým vyjádřením hustoty pravděpodobnosti je Gaussova křivka Obrázek 3.1, která je symetrická kolem střední hodnoty μ. Tvar křivky nám říká to, že nejčastěji při opakování náhodného pokusu budou vycházet hodnoty v okolí střední hodnoty. Funkce dosahuje maxima pro $x = \mu$ a odchylka σ nám udává v jaké vzdálenosti od střední hodnoty leží inflexní body.



Obrázek 3.2: Hustota pravděpodobnosti náhodné veličiny s rozdělením $N(\mu = 20; \sigma)$

Obrázek 3.2 znázorňuje, jak se se změnou směrodatné odchylky mění i tvar Gaussovy křivky. Čím nižší σ, tím je křivka vyšší a užší. Naopak s rostoucí směrodatnou odchylkou se křivka rozšiřuje a klesá. Je tomu tak, aby plocha pod křivkou zůstávala jednotková.



Obrázek 3.3: Hustota pravděpodobnosti náhodné veličiny s rozdělením $N(\mu; \sigma = 5)$

Co se děje při změně parametru μ si můžeme všimnout na obrázku Obrázek 3.3. Směrodatná odchylka nám udává maximum Gaussovy křivky. Při změně se nám tvar křivky nemění, pouze se na ose x posune.

Distribuční funkce náhodné veličiny X s normálním rozdělením má tvar

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt. \tag{3.2}$$

Poněvadž integrovanou funkci ve vztahu (3.2) není možné zapsat pomocí konečně mnoha elementárních funkcí, nelze vyjádřit distribuční funkci analyticky. Je možné vyjádřit distribuční funkci normální náhodné veličiny pomocí normovaného normálního rozdělení náhodné veličiny(Věta 3.1).

Charakteristická funkce s normálním rozdělením je ve tvaru

$$\psi(x) = e^{ix\mu - \frac{\sigma^2 x^2}{2}}. \tag{3.3}$$

3.1 Normované normální rozdělení

Normované (standardizované) normální rozdělení je speciálním případem normálního rozdělení, Obrázek 3.4. Jde o rozdělení náhodné veličiny $Z \sim N(0; 1)$, tudíž se jedná o

normální rozdělení se střední hodnotou $E(Z) = 0$ a roptylem $D(Z) = 1$. Zde je použito speciální značení funkcí, abychom je rozlišili od funkcí normálního rozdělení. *Hustota pravděpodobnosti* normovaného normálního rozdělení má tvar

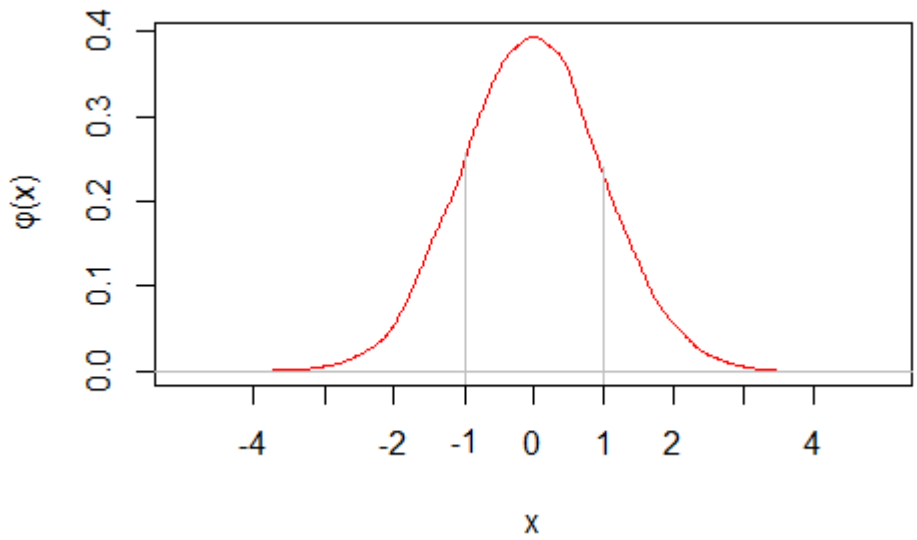
$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}, \text{pro} \quad z \in \mathbb{R}, \tag{3.4}$$

distribuční funkce je ve tvaru

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{z} e^{\frac{t^2}{2}} dt \tag{3.5}$$

a *charakteristická fuknce* je dána jako

$$\phi(x) = e^{-\frac{x^2}{2}}. \tag{3.6}$$



Obrázek 3.4: Hustota pravděpodobnosti normovaného normálního rozdělení

Věta 3.1 *Distribuční funkce náhodných veličin s normálním rozdělením s libovolnými parametry μ a σ lze počítat pomocí normovaného normálního rozdělení vztahem*

$$F_X(x) = \Phi(\frac{x - \mu}{\sigma}). \tag{3.7}$$

Důkaz. Nechť $X \sim N(\mu; \sigma^2)$ a $Z = \dfrac{X-\mu}{\sigma}$.

$Z = \dfrac{1}{\sigma}X - \frac{\mu}{\sigma}$, tj. náhodná veličina Z je lineární transformací náhodné veličiny X. Je zřejmé, že Z má rovněž normální rozdělení a $E(Z) = \dfrac{1}{\sigma} \cdot E(X) - \frac{\mu}{\sigma} = 0$, $D(Z) = \dfrac{1}{\sigma^2} \cdot D(X) = 1$, tj. $Z \sim N(0,1)$.

$$F_X(x) = P(X < x) = P(Z \cdot \sigma + \mu < x) = P(Z < \frac{x-\mu}{\sigma}) = \Phi(\frac{x-\mu}{\sigma})$$

■

Distribuční funkce normovaného normálního rozdělení je tabelována v tabulce v příloze D. V tabulce jsou uvedeny hodnoty distribuční funkce pouze pro $z > 0$. Ze symetrie hustoty pravděpodonosti $\varphi(x)$ je patrné, že je-li x záporné, pak jej určíme převodním vztahem doplňku

$$\Phi(-x) = 1 - \Phi(x) \tag{3.8}$$

Normální rozdělení také hraje důležitou roli při aproximaci některých náhodných veličin. Například pomocí limitních vět. Jako první si uveďme Lindebergovu-Lévyho větu.

Věta 3.1.1 (Lindebergova-Lévyho) *Nechť $X_1, X_2, ..., X_n$ jsou nezávislé náhodné veličiny se stejným (libovolným) rozdělením, stejnými středními hodnotami $EX_1 = EX_2 = ... = EX_n = \mu$ a se stejnými (konečnými) rozptyly $DX_1 = DX_2 = ... = DX_n = \sigma^2$. Potom náhodná veličina Y_n daná vztahem*

$$Y_n = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sigma\sqrt{n}}$$

má asymptoticky normované normální rozdělení $N(0,1)$.

Důkaz. Označme pro $k = 1, 2, \ldots$

$$C_k = X_k - \mu \text{ a } \psi_{C_k}(t) = E(e^{itC_k}).$$

Připomeňme si, že ψ_X je charakteristická funkce náhodné veličiny X definovaná jako $\psi_X(t) = E(e^{itX})$, kde $t \in \mathbb{R}$. Připomeňme také, že střední hodnota náhodné veličiny X je $E(X) = \mu$ a rozptyl $D(X) = \sigma^2$. Protože pro střední hodnotu platí $E(aX+b) = aE(X)+b$ a pro rozptyl $D(X) = E(X^2) - E(X)^2$, pak platí, že

$$E(C_k) = E(X_k - \mu) = E(X_k) - \mu = \mu - \mu = 0$$

$$D(C_k) = D(X_k) = \sigma^2.$$

Rozviňme charakteristickou funkci $\psi_{C_k}(t)$ pomocí Taylorova rozvoje, obecně: $\psi_{C_k}(t) = \sum_{j=0}^{n} \frac{(it)^j E(C_k^j)}{j!} + R_{n+1}(t)$, kde $R_{n+1}(t) = o(t^n)$ představuje chybu, tzn. $\lim_{t\to 0} \frac{R_{n+1}(t)}{t^n} = 0$. Protože máme zaručenu existenci prvních dvou momentů, pak pro $n = 2$ je Taylorův rozvoj $\psi_{C_k}(t) = 1 + \frac{itE(C_k)}{1!} + \frac{(it)^2 E(C_k^2)}{2!} + R_3(t) = 1 - \frac{\sigma^2 t^2}{2} + R_3(t)$, kde $\lim_{t\to 0} \frac{R_3(t)}{t^2} = 0$.

Položme

$$Z_k = \frac{C_k}{\sigma\sqrt{n}} = \frac{X_k - \mu}{\sigma\sqrt{n}}.$$

Protože platí

$$\psi_{a+bX}(t) = E(e^{(a+bX)it}) = E(e^{ita}e^{itbX}) = e^{ita}E(e^{(bt)iX}) = e^{ita}\psi_X(tb),$$

položíme–li $a = 0$ a $b = \frac{1}{\sigma\sqrt{n}}$, můžeme pro $\psi_{Z_k}(t)$ psát

$$\psi_{Z_k}(t) = \psi_{C_k}\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 - \frac{\sigma^2 t^2}{2\sigma^2 n} + R_3\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + R_3\left(\frac{t}{\sigma\sqrt{n}}\right)$$

a přitom

$$\lim_{t\to 0} \frac{R_3\left(\frac{t}{\sigma\sqrt{n}}\right)}{\frac{t^2}{\sigma^2 n}} = 0 \quad \Leftrightarrow \quad \text{pro pevné } t \in \mathbb{R} \quad \frac{\sigma^2}{t^2} \lim_{n\to\infty} nR_3\left(\frac{t}{\sigma\sqrt{n}}\right) = 0.$$

Nakonec položme

$$Y_n = Z_1 + \cdots + Z_n = \frac{\sum_{q=1}^{n}(X_q - \mu)}{\sigma\sqrt{n}} = \frac{\sum_{q=1}^{n} X_q - n\mu}{\sigma\sqrt{n}}.$$

Protože jde o součet nezávislých náhodných veličin, pak pro jejich charakteristické funkce platí

$$\begin{aligned}\psi_{Y_n}(t) = \psi_{Z_1+\cdots+Z_n}(t) &\overset{nez.}{=} \prod_{k=1}^{n} \psi_{Z_k}(t) = \prod_{k=1}^{n} \psi_{C_k}\left(\tfrac{t}{\sigma\sqrt{n}}\right) = \left[\psi_{C_k}\left(\tfrac{t}{\sigma\sqrt{n}}\right)\right]^n \\ &= \left[1 - \tfrac{t^2}{2n} + R_3\left(\tfrac{t}{\sigma\sqrt{n}}\right)\right]^n\end{aligned}$$

Počítejme limitu

$$\begin{aligned}\lim_{n\to\infty} \psi_{Y_n}(t) &= \lim_{n\to\infty} \left[1 - \tfrac{t^2}{2n} + R_3\left(\tfrac{t}{\sigma\sqrt{n}}\right)\right]^n = \lim_{n\to\infty} \left[1 - \frac{\frac{t^2}{2} + nR_3\left(\frac{t}{\sigma\sqrt{n}}\right)}{n}\right]^n \\ &= \lim_{n\to\infty} \left[1 - \frac{\frac{t^2}{2}}{n}\right]^n = e^{-\frac{t^2}{2}}\end{aligned}$$

což je charakteristická funkce $N(0,1)$. Pokud jde tedy veličina Y_n svou charakteristickou funkcí k charakteristické funkci normovaného normálního rozdělení, pak musí ke stejnému rozdělení jít i svou hustotou pravděpodobnosti a distribuční funkcí, tedy $f_Y(u) \to \varphi(u)$ a $F_Y(u) \to \Phi(u)$. [2] ∎

Z věty Věta 3.1.1 plyne, že pro dostatečně velká n, lze rozdělení náhodné veličiny $X = \sum_{i=1}^{n}(X_i)$ aproximovat rozdělením $N(n\mu, n\sigma^2)$, tj. X má asymptoticky normální rozdělení, stejně tak veličina $\overline{X} = \frac{\sum_{i=1}^{n}(X_i)}{n}$ má asymptoticky rozdělení $N(\mu, \frac{\sigma^2}{n})$.

Pak také existuje Moivreova-Laplaceova věta, která vyjadřuje konvergenci binomického rozdělení k normálnímu rozdělení.

Věta 3.1.2 (Moivreova-Laplaceova) *Nechť $X \sim Bi(n, \pi)$; $EX = n\pi$; $DX = n\pi(1-\pi)$. Potom pro velká n platí, že $X \approx N(n\pi; n\pi(1-\pi))$.*

Důkaz. Binomická náhodná veličina $X \sim Bi(n, \pi)$ je součtem n nezávislých náhodných veličin s alternativním rozdělením. Označme $X_i \sim A(\pi)$. Pak posloupnost náhodných veličin $\{X_i\}_{i=1}^{\infty}$ s konečnou střední hodnotou $E(X_i) = \pi$ a konečným rozptylem $D(X_i) = \pi(1-\pi)$ splňuje Lindebergovu–Lévyho větu, takže platí

$$X = \sum_{i=1}^{n} X_i \sim N(n \cdot E(X_i); n \cdot D(X_i))$$

$$X = \sum_{i=1}^{n} X_i \sim N(n\pi; n\pi(1-\pi)$$

Tím je věta dokázána. [2]. ∎

4 Grafické metody pro ověření normality dat

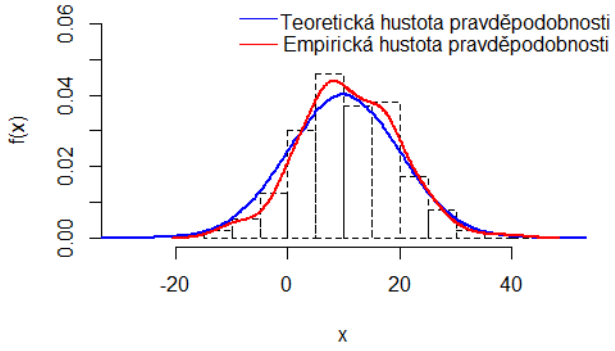
Normalita dat je předpoklad pro mnoho testů hypotéz a intervalových odhadů. Protože má normální rozdělení dobré vlastnosti, můžeme jejich zásluhou o vyšetřovaném souboru vyvodit jisté závěry. Snažíme se proto u zkoumaných jevů zjistit, zda pocházejí z normálního rozdělení. Pokud bychom používali parametrické testy(tj. testy vyžadující normalitu dat) na jiné rozdělení, došli bychom k nepřesným výsledkům. Normalitu dat můžeme ověřit několika způsoby, jak graficky, tak různými testy normality. Nejpoužívanějším testem pro zjištění normality jsou Shapirův-Wilkův test nebo Kolmogorovův-Smirnovův test.

Grafické metody nejsou sice tak přesné jako exaktní, ale mohou nám sdělit informaci o rozdělení dat, konkrétně zda data pochází z normálního rozdělení. Jednoduchý způsob, jak přibližně odhadnout, zda data mají normální rozdělení, je použití P-P grafu, Q-Q grafu nebo N-P grafu. U P-P grafů se porovnávají pravděpodobnosti. Q-Q grafy jsou grafy založené na porovnávání kvantilů naměřených s kvantily teoretickými. Bohužel neposkytují objektivní důvod k potvrzení nebo zamítnutí nulové testové hypotézy, která nám říká, že data pochází z normálního rozdělení. Grafické metody, na rozdíl od některých testů nekladou omezení na rozsah výběru.

4.1 Srovnání empirické a teoretické hustoty

V teorii pravděpodobnosti je odvozena spousta teoretických modelů rozdělení pravděpodobnosti. Těmto modelům se mohou empirická rozdělení více či méně přibližovat. Tudíž všechny empirické křivky hustoty pravděpodobnosti pro možné výběry se pohybují přibližně okolo teoretické hustoty pravděpodobnosti. Empirické charakteristiky představují náhodné veličiny, protože se mění od jednoho náhodného výběru ke druhému náhodnému výběru, zatímco teoretické charakteristiky základního souboru představují vždy určité číslo. Čím bude větší rozsah výběrového statistického souboru, tím bude empirická křivka blíže skutečnému tvaru teoretické křivky.

Pomocí grafické analýzy můžeme metodou srovnání se standardními modely pouze odhadnout typ rozdělení, nikoli objektivní míru shody dat s teoretickým modelem.



Obrázek 4.1: Porovnání empirické a teoretické hustoty pravděpodobnosti

Všimněme si, že na obrázku Obrázek 4.1 máme příklad grafického vyjádření rozdělení spojité náhodné veličiny pomocí teoretické křivky ve vztahu k empirické křivce rozdělení. Vykreslení teoretické hustoty odpvídá hladkému průběhu, oproti tomu empirická hustota je jakožto aproximace tvořená reálnými daty.

4.2 Kvantilově-kvantilový graf (Q-Q graf)

Q-Q graf je jedním z nejvyužívanějších grafů pro hodnocení normality. Tato grafická metoda spočívá v nanesení empirických kvantilů a teoretických kvantilů rozdělení pravděpodobnosti proti sobě.

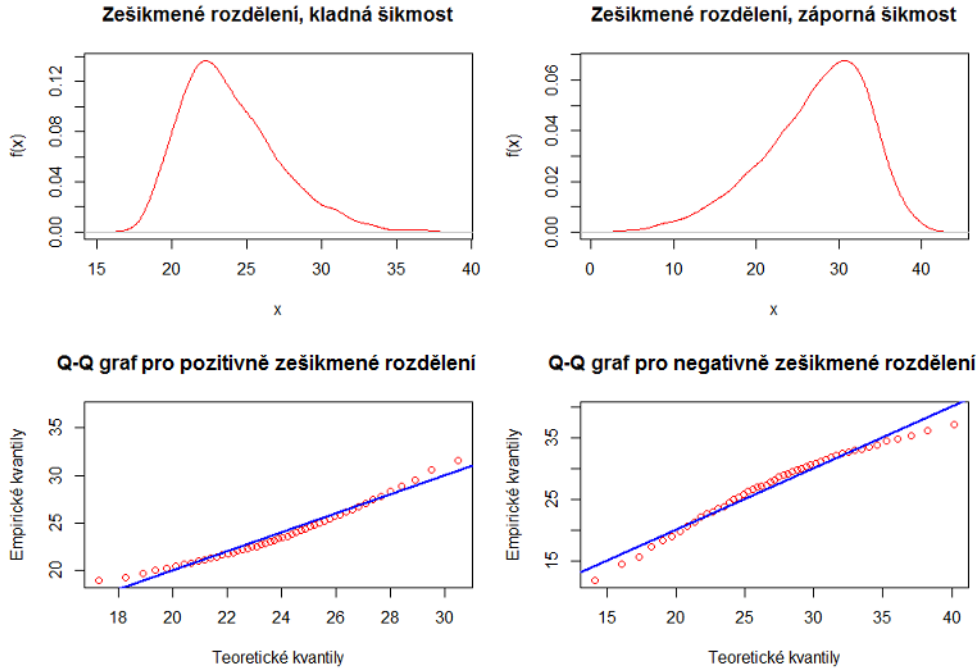
Konstrukce grafu spočívá v tom, že na vodorovné ose jsou teoretické kvantily normálního rozdělení a na ose svislé jsou empirické kvantily zjištěné z výběru.

Tudíž na osu x kvantily x_{α_j} vybraného rozdělení, kde $\alpha_j = \dfrac{j - r_{\alpha dj}}{n + n_{\alpha dj}}$, kde $r_{\alpha dj}$ a $n_{\alpha dj}$ jsou korigující faktory, které jsou $\leq 0,5$. Korigující faktory implicitně volíme $r_{\alpha dj} = 0,375$ a $n_{\alpha dj} = 0,25$. Graf, kde za korigující faktory dosadíme hodnoty $r_{\alpha dj} = 0,3175$, $n_{\alpha dj} = 0,365$, bývá někdy označován jako normální pravděpodobnostní graf (N-P graf). Jsou-li nějaké uspořádané hodnoty stejné, potom za j bereme průměrné pořadí odpovídající stejné skupině hodnot.

Na osu y vynášíme uspořádané hodnoty $x_{(j)}$. Tyto hodnoty představuje $q_j\%$ kvantil, kde $q_i = 100(j - 0.5)/n$.

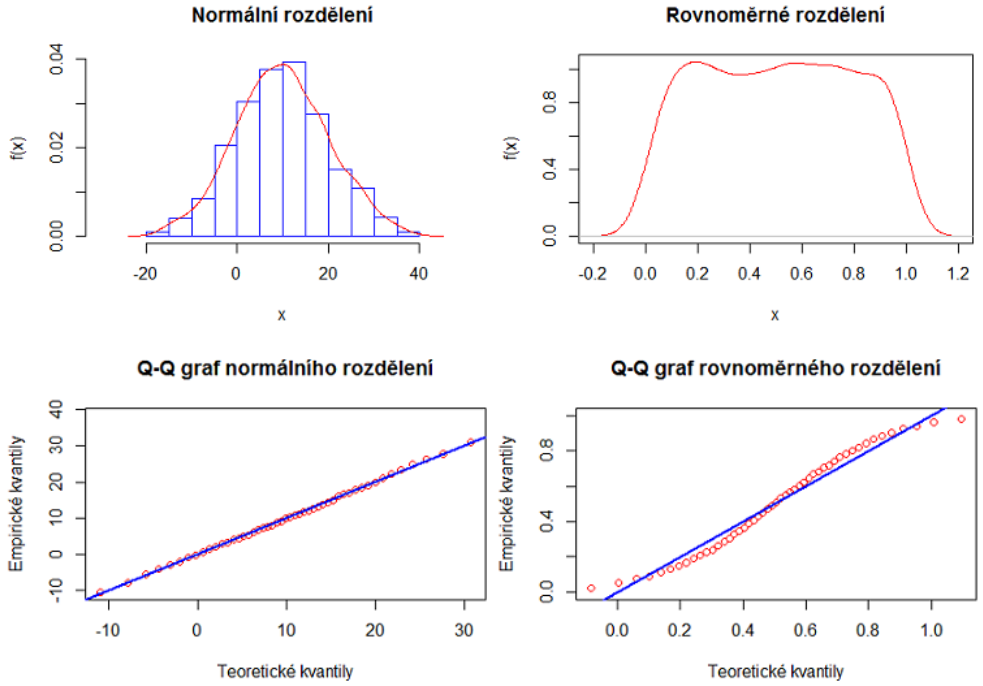
$$x_{(j)} \approx x_{\alpha_j}$$

Poté se body $[x_{\alpha_j}, x_{(j)}]$ proloží přímka, která je osou 1. a 3. kvadrantu.



Obrázek 4.2: Grafické zobrazení zešikmeného rozdělení pomocí Q-Q grafů

Obrázek 4.2 pro grafické znázornění zešikmeného rozdělení, na kterém můžeme vidět, že data v grafu jsou zešikmená, neleží na ose prvního a třetího kvadrantu. Q-Q graf s pozitivně zešikmenými daty tvoří konvexní křivku, nikoli přímku. Stejně tomu je i u Q-Q grafu pro negativní zešikmení, který má křivku spíš konkávní. Díky tomu můžeme usuzovat, že data pravděpodobně nepocházejí z normálního rozdělení.



Obrázek 4.3: Grafické zobrazení Q-Q grafů pro náhodné výběry z normálního a rovnoměrného rozdělení

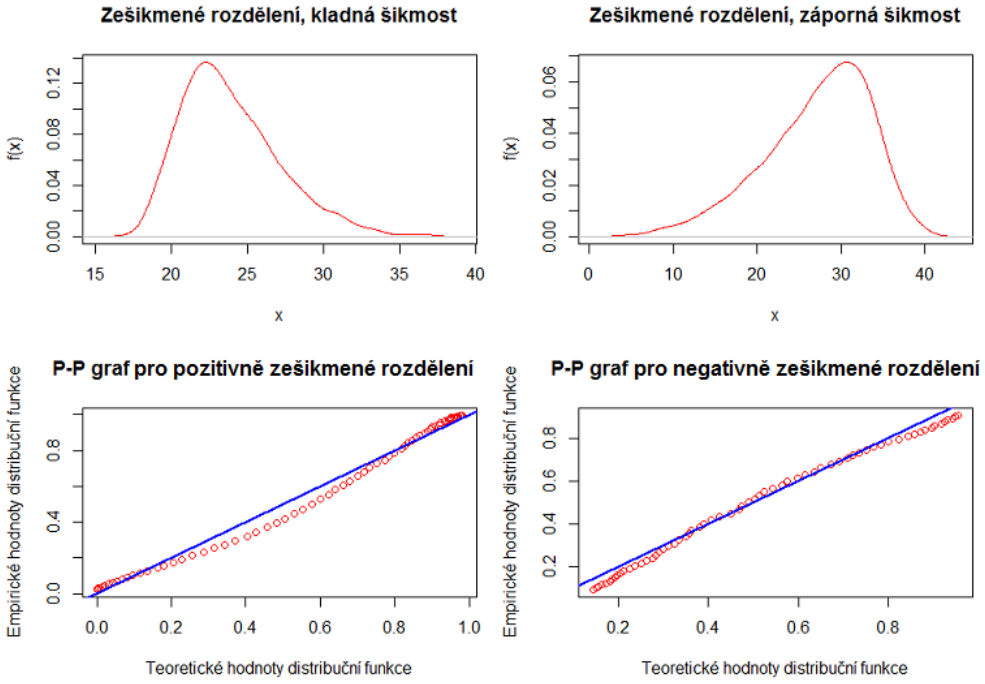
Obrázek 4.3 je rozdělen na dvě části. Vlevo je vidět, že data můžeme považovat za data pocházející z normálního rozdělení, protože nemůžeme z Q-Q grafu zřetelně rozlišit odchylku od přímky. Měli bychom se ale ještě přesvědčit testem normality. V pravé části obrázku vidíme původ rovnoměrného rozdělení. Příčinou může být úmyslné zkreslení dat odstraněním příliš nízkých a vysokých hodnot. Na Q-Q grafu lze vidět, že body na přímce neleží.

4.3 Pravděpodobnostní graf (P-P graf)

Můžeme jej nalézt také pod názvem procentní graf. P-P graf je jedna námi zmiňovaná grafická metoda, používá se podobně jako Q-Q graf, ale je zkonstruovaný jinak.

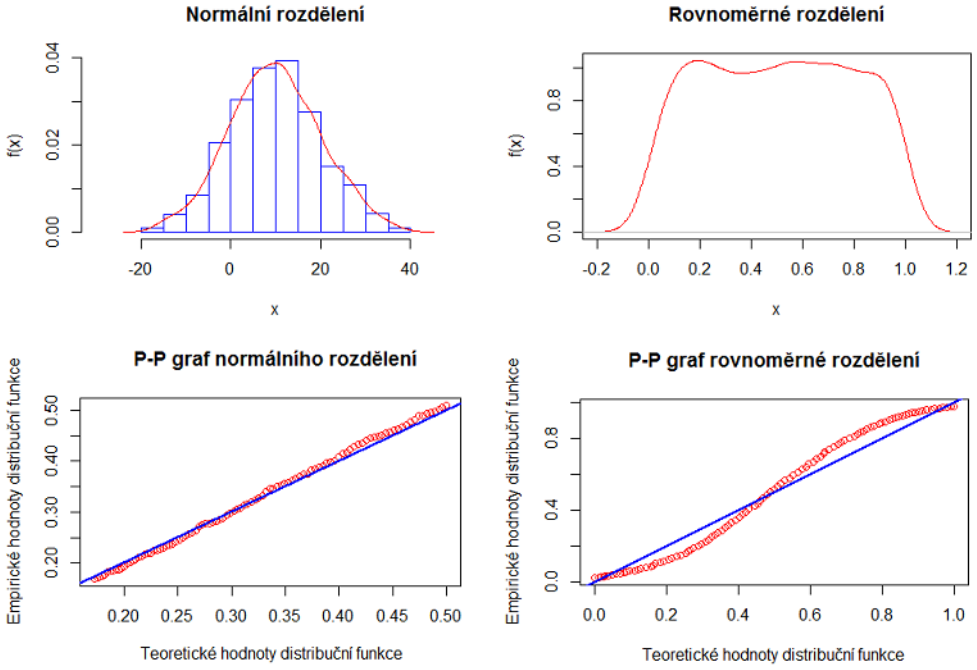
Konstrukce P-P grafu spočívá v tom, že na jednu osu vyneseme hodnoty empirické distribuční funkce, na osu druhou hodnoty teoretické distribuční funkce. Nejprve si spočítáme normované hodnoty $z_{(j)} = \frac{x_{(j)} - \overline{x}}{s}$, kde $\overline{x}$ je aritmetický průměr hodnot $x_1, ..., x_n$ a s je jejich směrodatná odchylka. Na vodorovnou osu x vyneseme $\Phi(z_{(j)})$ a na svislou

osu y hodnoty $F(z_{(j)}) = \frac{j}{n}$. Jsou-li některé honoty $x_{(1)} \leq ... \leq x_{(n)}$ identické, pak za j bereme jejich průměrné pořadí odpovídající stejné skupině hodnot. Pokud výběr pochází z normálního rozdělení, měly by body $[\Phi(z_{(j)}, F(z_{(j)}]$ vytvořit úsečku, tentokrát ale s krajními body [0,0],[1,1].



Obrázek 4.4: Grafické zobrazení zešikmeného rozdělení pomocí P-P grafů

Vykreslení P-P grafů dopadlo obdobně jako u předchozích grafických metod. Obrázek 4.4 je znázorňením P-P grafů pro zešikmená data. Opět vidíme, že hodnoty neleží na přímce, proto předpokládáme, že data pocházejí z jiného než normálního rozdělení.



Obrázek 4.5: P-P grafy pro náhodné výběry z normálního a rovnoměrného rozdělení

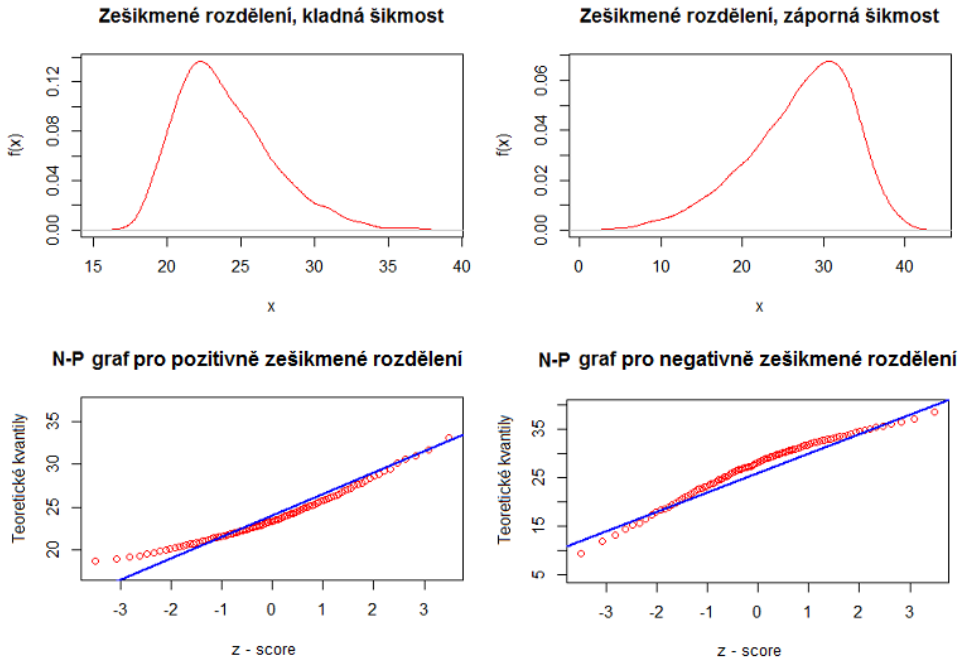
Můžeme si všimnout, že na obrázku Obrázek 4.5 napravo máme P-P graf pro data pocházející z normálního rozdělení. Hodnoty vynesené v grafu neleží přímo na přímce, ale leží velice blízko přímky, proto nemůžeme zamítnout, že data pocházejí z normálního rozdělení. Vlevo na obrázku je jasně vidět, že data neleží na přímce.

4.4 Normální pravděpodobnostní graf (N-P graf)

N-P graf je poslední námi zmiňovaný způsob grafických metod, kterým dokážeme posoudit, zda data pocházejí z normálního rozdělení a ověřit tak normalitu dat.

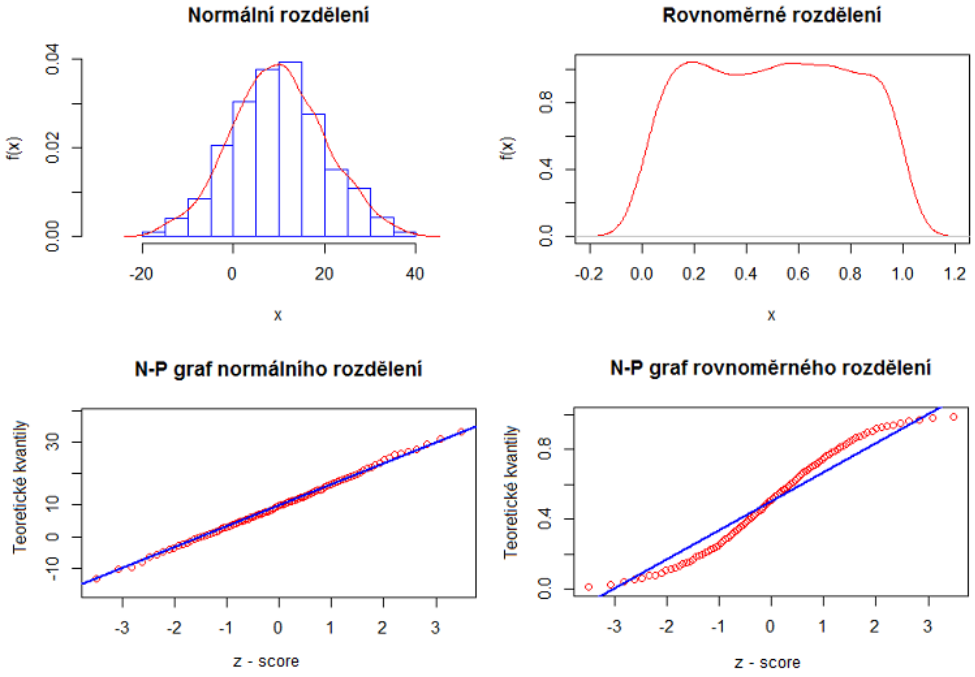
Graf se konstruuje tak, že na vodorovnou osu x se vynáší hodnoty z-score, tj. $\frac{x_{(i)} - \overline{x}}{s}$, na svislou osu y kvantily u_{α_j}, kde $\alpha_j = \frac{3j-1}{3n+1}$ a kde j je pořadí j-té uspořádané hodnoty.

Pokud tyto data pochází z normálního rozdělení, pak body $[x_{(j)}, u_{\alpha_j}]$ v grafu leží na přímce. Jakékoli odbočení od této přímky vyjadřuje odchylku od normálního rozdělení.



Obrázek 4.6: Grafické zobrazení zešikmeného rozdělení pomocí N-P grafů

Obrázek 4.6 znázorňuje zešikmená data. Data vpravo se řadí do pomyslné konkávní křivky, což znamená, že jsou kladně zešikmená. Nalevo tomu je přesně naopak, data se řadí do konvexní křivky a proto jsou záporně zešikmená. Data tudíž nepocházejí z normálního rozdělení.



Obrázek 4.7: Grafické zobrazení N-P grafů z náhodných výběrů normálního a rovnoměrného rozdělení

Obrázek 4.7 znázorňuje dva typy rozdělení dat, které mají původ z normálního a rovnoměrného rozdělení. Z N-P grafu pro normální rozdělení jde krásně vidět, že body leží téměř na přímce, proto data splňují vlastnosti normality. Rovnoměrné rozdělění tvoří v N-P grafu body, které neleží na přímce.

5 Testy normality

Už jsme si ukázali, jak nám grafické metody mohou pomoci při určení, zda se data podobají normálnímu rozdělení, přesto ale potřebujeme testy, které nám mohou poskytnout informaci, zda můžeme či nemůžeme zamítnout nulovou hypotézu o normalitě. Testy normality dat tedy probíhají tak, že na základě hodnoty testové statistiky nezamítáme popřípadě zamítáme hypotézu o tom, zda data pocházejí z normálního rozdělení.

5.1 Chí-kvadrát test dobré shody

Je pravděpodobně nejstarším testem pro shodu rozdělení výběru (tzv. empirického rozdělení) s hypotetickým rozdělením (tzv. teoretickým rozdělením), tedy i testem normality. Později se objevilo více kvalitnějších testů normality, proto se tento test stal ne tolik používaným testem.

Chí-kvadrát test dobré shody je založen na posouzení rozdílu mezi skutečnými četnostmi výskytu hodnot a očekávanými četnostmi, odpovídající příslušnému předpokládanému normálnímu rozdělení.

Tento test použijeme nejčastěji ve dvou vyskytujících se situacích. První situací je ta, kdy nám nulová hypotéza udává typ rozdělení, které testujeme, ale také jeho parametry, mluvíme o úplně specifikovaném modelu. Druhá varianta nastane v případě, že nám H_0 udává typ rozdělení, ale nejsou zadány všechny parametry. Tehdy hovoříme o neúplně specifikovaném modelu. V našem případě testujeme variantu druhou, kdy chceme otestovat shodu našeho empirického rozdělení výběru s rozdělením teoretickým, jehož všechny parametry neznáme, musíme je odhadnout. Za střední hodnotu bereme průměr $\overline{x}$ a místo směrodatné odchylky použijeme výběrovou odchylku s. Počet odhadovaných parametrů označujeme písmenem c.

Nejprve si musíme náš datový soubor rozdělit do několika intervalů, potom při pozorování každého intervalu musíme zjistit četnosti $\{n_i\}$ jednotlivých kategorií (intervalů), kde $i = 1, 2, ..., k$. Dále musíme zjistit pravděpodobnosti $\{p_i\}$, že hodnoty rozdělení, se kterým chceme data srovnat, spadají do i-tého itervalu. Normální distribuční funkci si označíme symbolem $F_0(x)$ a symbolem $F(x)$ označíme distribuční funkci, kterou náhodná veličina skutečně má.

Potom za nulovou hypotézu H_0 považujeme tvrzení $F(x) = F_0(x)$ oproti tomu alternativa H_1 je dána $F(x) \neq F_0(x)$.

Test rozhoduje, zda je rozdíl mezi distribucemi způsobený náhodně a datový soubor pochází z normálního rozdělení, nebo je rozdíl velký a výběr nepochází z odpovídajícího rozdělení, ale z jiného. Tento rozdíl mezi četnostmi zachycuje testovací statistika, která je

dána tvarem

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}, \tag{5.1}$$

kde k je počet slučitelných intervalů náhodné veličiny X. Jako pozorovanou četnost značíme n_i, pro každé $i = 1, 2, ..., k$. Za teoretickou četnost považujeme np_i, pro $i = 1, 2, ..., k$, vypočítanou za předpokladu platnosti H_0, přičemž n je rozsah výběru.

Kritická hodnota je stanovena pro hladinu významnosti α. Hladina významnosti se obvykle volí $\alpha = 0,05$ nebo $\alpha = 0,01$, my používáme $\alpha = 0,05$. Testová statistika má χ^2 rozdělení s $k - c - 1$ stupni volnosti. Kritická hodnota $\chi^2_{k-c-1}(\alpha)$ je α kvantil tohoto rozdělení. Nulovou hypotézu H_0 na hladině významnosti α zamítáme v případě je-li testová statistika větší nebo rovna kritické hodnotě, kde c označuje počet odhadovaných parametrů rozdělení náhodné veličiny X. P-hodnotu určíme jako $1 - F(x_{OBS})$, kde x_{OBS} je pozorovaná hodnota testové statistiky.

5.2 Lillieforsův (Kolmogorovův-Smirnovův) test

Lillieforsova varianta je obdobou Kolmogorova-Smirnovova testu pro neúplně specifikované testy. Nevyžaduje přesnou znalost parametrů μ a σ, ale stačí jejich odhady.

Testujeme, zda náhodná veličina má předpokládané teoretické rozdělení. V našem případě, zda má náhodná veličina normální rozdělení pravděpodobnosti. Test se pro výběry o rozsahu $n \geq 100$ stal předepisován normou ČSN 01 0225 [7] díky dobrým vypovídacím schopnostem.

Testujeme nulovou hypotézu H_0, zda data odpovídají normálnímu rozdělení, pokud rozdíly mezi empirickou distribuční funkcí a teoretickou distibuční funkcí jsou statisticky nevýznamné. Alternativní hypotézou H_1 je opak nulové hypotézy (data pocházejí z jiného než normálního rozdělení).

Našich n hodnot výběru $x_{(1)}, \cdots, x_{(n)}$ odpovídá uspořádaným hodnotám výběru. Funkce $F_0(x)$ je teoretickou distribuční funkcí. Potom je důležité popsat rozdělení testované náhodné veličiny. To bude představovat výběrová distribuční funkce $F_n(x)$. Distribuční funkce $F_n(x)$ představuje odhad skutečného rozdělení a je ve tvaru

$$F_n(x) = \begin{cases} 0, & \text{pro} \quad x < x_1 \\ \frac{i}{n}, & \text{pro} \quad x_i < x < x_{i-1}, \quad \text{kde} \qquad i = 1, \cdots, n-1 \\ 1, & \text{pro} \quad x \geq x_n. \end{cases} \tag{5.2}$$

Testová statistika Kolmogorova-Smirnovova testu je tedy definována vztahem

$$D_n = \sup|F_n(x) - F_0(x)|, \tag{5.3}$$

kde D_n je maximální odchylka empirické a teoretické distribuční funkce a $D_n(\alpha)$ je kritická hodnota, α je hladina významnosti.

Pokud je hodnota testové statistiky D_n větší než hodnota kritická $D_n(\alpha)$, hypotéza H_0 je zamítnuta na hladině významnosti α.

Kritická hodnota závisí na hladině významnosti α a rozsahu výběru n. Pro výběry o rozsahu $n < 100$ jsou kritické hodnoty v např. [8]. Pro $n > 100$ je možné použít aproximaci. Přibližné výrazy jsou uvedeny v [3] a jsou ve tvaru

$$D_n(\alpha) = \sqrt{\frac{1}{2n} \cdot ln(\frac{2}{\alpha})}. \tag{5.4}$$

Pro Lillieforsův test existuje v programu RStudio funkce `lillie.test(x)`, kterou najdeme v balíčku `nortest`, kde `x` je vstupní vektor testovaných dat. Kritickou hodnotu jsme dopočítali ručně. Tato hodnota bude pro každé rozdělení stejná, protože máme pro každé rozdělení použitý stejný počet testovaných dat.

5.3 Testy založené na šikmosti a špičatosti

Uvedeme si testy normality, které jsou založeny na výběrové šikmosti a špičatosti, nebo na jejich kombinaci. Při těchto testech vycházíme z předpokladu, že je-li výběr z nějakého rozdělení pravděpodobnosti, potom pro jeho šikmost a_3 a špičatost a_4 platí, že mají asymptoticky normální rozdělení s parametry

$$E(a_3) = 0, \qquad E(a_4) = 3 - \frac{6}{n+1} \tag{5.5}$$

$$D(a_3) = \frac{6(n-2)}{(n+1)(n+3)}, \qquad D(a_4) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}. \tag{5.6}$$

5.3.1 Test založený na šikmosti

V testu založeném na výběrové šikmosti testujeme nulovou hypotézu H_0, která nám říká, že se jedná o normální rozdělení. Oproti ní alternativní hypotéza H_1 nám říká, že výběr pochází z nějakého jiného rozdělení, které je nesymetrické.

Pro malé rozsahy výběru $n > 25$ je kritická hodnota statistiky a_3 uvedena v tabulkách [1]. Pro větší rozsah výběrů než $n > 200$ pro šikmost a_3, musíme použít aproximaci normálním rozdělením, které vychází z centrální limitní věty. Testovou statistiku počítáme tedy s tím, že náhodné veličiny

$$U_3 = \frac{a_3}{\sqrt{D(a_3)}}, \tag{5.7}$$

mají normované normální rozdělení.

Na hladině významnosti α, potom zamítáme nulovou hypotézu v případě, je-li $|U_3| > u_{1-\frac{\alpha}{2}}$, kde u_α je α-kvantil normovaného normálního rozdělení $N(0,1)$.

5.3.2 Test založený na špičatosti

Stejně jako u testu šikmosti, tak pro test založený na výběrové špičatosti, uvažujeme nulovou a alternativní hypotézu. Zde je H_0 o normalitě oproti alternativě H_1, že výběr pochází z rozdělení, které je jiné než normální a výběr se liší špičatostí.

Kritické hodnoty závisí na velikosti našeho rozsahu výběru. Pro malé rozsahy $n > 50$ je kritická hodnota špičatosti a_4 uvedena v tabulkách [1]. V našem případě je $n = 1000$, proto použijeme opět aproximaci normálním rozdělením vycházející z centrální limitní věty, pro výběry o rozsahu $n > 500$. Testová statistika má potom tvar

$$U_4 = \frac{a_4 - E(a_4)}{\sqrt{D(a_4)}} \tag{5.8}$$

a náhodné veličiny mají normované normální rozdělení.

Alternativní hypotézu přijímáme pokud $|U_4| > u_{1-\frac{\alpha}{2}}$, kde u_α je α-kvantil normovaného normálního rozdělení $N(0,1)$. Symbolem α značíme hladinu významnosti.

5.3.3 Test založený na šikmosti i špičatosti

U tohoto testu zkoumáme nulovou hypotézu H_0, jestliže výběr dat pochází z normálního rozdělení pravděpodobnosti a alternativní hypotézu H_1, která nám říká, že data pochází z jiného rozdělení pravděpodobnosti než normálního a odlišuje se šikmostí a špičatostí.

Pro výběry o rozsahu $n > 200$ použijeme testové statistiky šikmosti (5.7) a špičatosti (5.8). Potom je test založen na náhodné veličině

$$U_3^2 + U_4^2 = \frac{a_3^2}{D(a_3)} + (\frac{(a_4 - E(a_4))^2}{D(a_4)}) \sim \chi_2^2, \tag{5.9}$$

které má rozdělení χ^2 o dvou stupních volnosti.

Na hladině významnosti $\alpha = 0,05$ zamítáme hypotézu o normalitě, pokud je $U_3^2 + U_4^2 \geq \chi_2^2(\alpha)$.

Abychom dále v kapitole 6 Srovnání testů normality, mohli srovnávat testy i s testem založeným na výběrové šikmosti i špičatosti, potřebovali jsme si naprogramovat funkci, která nám spočítá a uloží výsledek testové statistiky, tuto fukci naleznete v příloze A.

5.4 Shapirův-Wilkův test

Shapirův-Wilkův test je založený na principech regresní analýzy. Jedná se o nejpoužívanější test pro malé až střední rozsahy dat. Výběry mohou být dokonce až o velikosti

$n \leq 50$ to je předepisováno i normou ČSN 01 0225 [7]. Testová statistika je definována vztahem

$$W = \frac{(\sum_{i=1}^{n} a_i \cdot x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{5.10}$$

kde n je rozsah výběru, $\overline{x}$ je výběrový průměr, $x_{(i)}$ označují řádkové statistiky a a_i jsou váhy, odvozovány ze středních hodnot a variační matice pořadových statistik prostého náhodného výběru z $N(0,1)$ rozsahu n.

V případě nulové hypotézy H_0 předpokládáme, že data pochází z normálního rozdělení. Za hypotézu alternativní H_1 považujeme opak nulové hypotézy.

Pro Shapirovův-Wilkův test normality dat jsem použila v programu Rstudio funkci `shapiro.test(x)`, která je v programu již definována. Test posuzujeme na hladině významnosti $\alpha = 0,05$. Vstupním argumentem `x` myslíme testovaný vektor dat.

5.5 Andersonův-Darlingův test

Jedná se o test založený na analýze empirické distribuční funkce testovaného datového souboru.

Pro Andersonův-Darlingův test uvažujeme nulovou hypotézu $H_0 : F_n(x) = F_0(x)$, kde $F_n(x)$ označuje empirickou distribuční funkci a $F_0(x)$ je distribuční funkce normálního rozdělení. Proti tomu uvažujeme alternativní hypotézu $H_1 : F_n(x) \neq F_0(x)$.

Testová statistika Andersonova-Darlingova testu pro ověřování normality n-prvkového výběru je definovaná vztahem

$$AD = -\frac{1}{n} \cdot \sum_{j=1}^{n}(2i-1) \cdot (lnz_i + ln(1 - z_{n-i+1})) - n, \tag{5.11}$$

kde z_i jsou hodnoty distribuční funkce normovaného normálního rozdělení dány vztahem

$$z_i = \Phi(\frac{x_{(i)} - \overline{x}}{s}). \tag{5.12}$$

Hodnota $\overline{x}$ je průměr a s je směrodatná odchylka.

Hypotéza H_0 se zamítá na hladině významnosti α, pokud je hodnota testové statistiky AD větší než kritická hodnota. Pro dost velký rozsah n výběru je přibližná hodnota $0,95$ kvantilu rovna přibližně

$$AD_{0,95} = 1,0348 \cdot (1 - \frac{1,013}{n} \breve{} \frac{0,93}{n^2}). \quad [3] \tag{5.13}$$

Pro otestování Andersonova-Darlingova testu normality jsem použila program Rstudio. V programu je předdefinovaná fuknce `ad.test(x)`, která se nachází v balíčku `nortest`. Test používá hladinu významnosti $\alpha = 0,05$. Vstupním argumentem `x` myslíme testovaný vektor dat.

6 Srovnání testů normality

Srovnání provádíme na výše zmíněných testech normality, které si aplikujeme na náhodné výběry ze základních rozdělení pravděpodobnosti. V našem případě na výběry z normálního, rovnoměrného, exponenciálního, logaritmicko-normálního a studentova rozdělení.

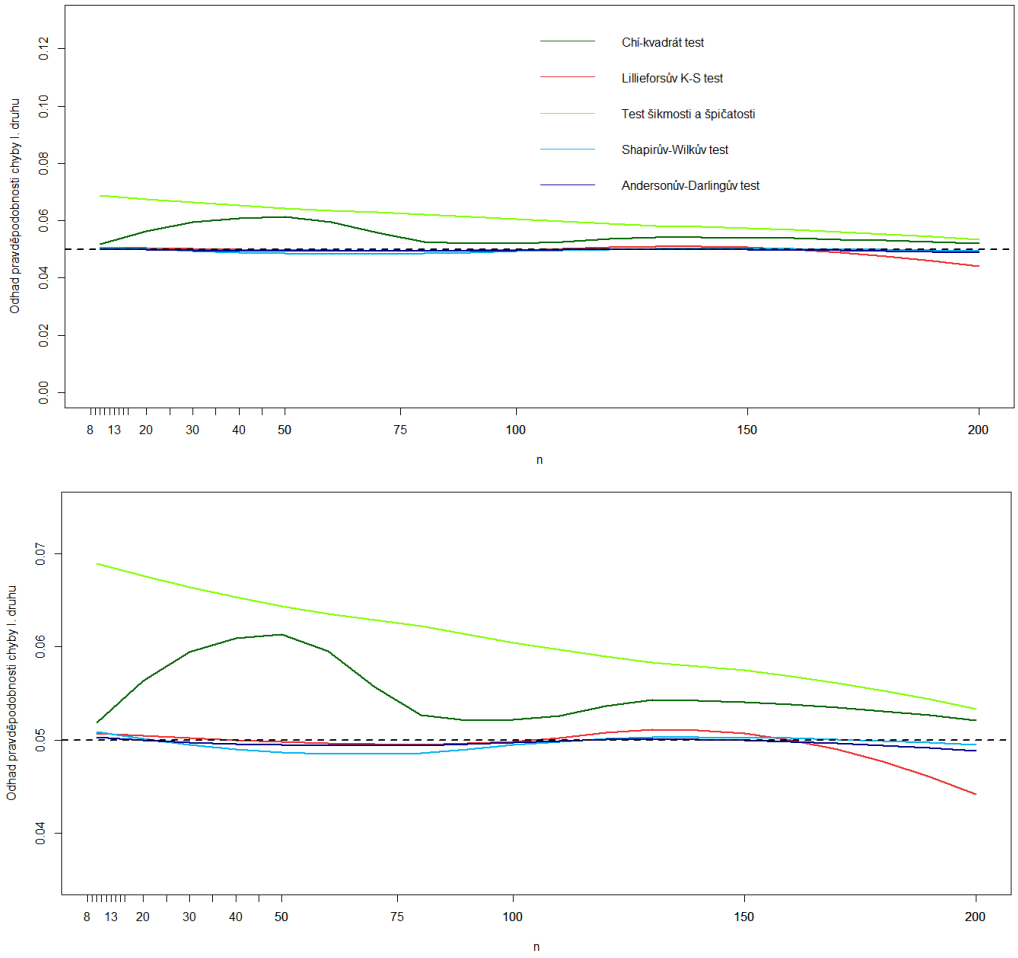
Jen pro připomenutí si uveďme, jaké chyby mohou nastat.

- *Chyba I. druhu* nastane, když zamítneme normalitu u dat, která jsou výběrem z normálního rozdělení pravděpodobnosti. Pravděpodobnost chyby I. druhu se značí písmenem α a bývá nazývaná hladinu významnosti.
- *Chyba II. druhu* nastane, pokud nezamítneme normalitu u dat, která **nejsou** výběrem z normálního rozdělení. Pravděpodobnosti chyby II. druhu značíme jako β.

Srovnání lze získat tak, že se testu předloží k ověření daný počet výběrových vektorů s určitými vlastnostmi. Ověřujeme tedy nulovou hypotézu o normalitě, ze kterého výběr dat pochází. Jelikož je předem známo, zda výběr z normálního rozdělení pravděpodobnosti pochází nebo ne, je možné určit, v kolika procentech případů došlo k chybě I. nebo II. druhu.

6.1 Srovnání podle chyby I. druhu

Při porovnávání testů normality podle chyby I. druhu jsme postupovali tak, že jsme v programu RStudio prováděli testy nad náhodně vygenerovanými daty. Pro každý rozsah výběrů bylo generováno 50 000 realizací náhodného výběru z normálního rozdělení pravděpodobnosti a pro každou u nich byla provedena sekvence všech srovnávaných testů na hladině významnosti 0,05. Odhad skutečnéptavděpodobnosti chyby I. druhu byl stanoven jako relativní četnost testů, v nichž došlo k neoprávněnému zamítnutí hypotézy o normalitě dat.



Obrázek 6.1: Odhad pravděpodobnosti chyby I. druhu při rozhodování nulové hypotézy o normalitě dat

Obrázek 6.1 zubrazuje dva grafy, kde vykreslujeme odhady praděpodobností chyby prvního druhu vzhledem k rozsahu výběru. Na spodním obrázku vidíme stejný graf jako na horním, který je pouze přiblížený.

Z grafu plyne, že pravděpodobnost chyby u Shapirova-Wilkova i Andersonova-Darlingova testu normality se ustaluje až, když se zvyšuje rozsah výběru. Lillieforsova varianta Kolmogorova-Smirnovova testu má pravděpodobnost chyby nezávislou na velikosti výběrového vektoru. Nejhůře dopadl Chí-kvadrát test dobré shody a test založený na výběrové šikmosti i špičatosti, který se se svojí pravděpodobností chyby pohybuje nad hladinou významnosti $\alpha = 0.05$.

Abychom se o odhadech pravděpodobnosti dozvěděli víc, má smysl podívat se i na výběrové charakteristiky.

Odhad pravděpodobnosti chyby I. druhu	CHI	KS	SS	SW	AD
Míry polohy					
minimum	0.0322	0.0447	0.0530	0.0466	0.0483
dolní kvartil	0.0492	0.0493	0.0566	0.0492	0.0491
medián	0.0529	0.0502	0.0595	0.0496	0.0497
průměr	0.0555	0.0495	0.0605	0.0496	0.0497
horní kvartil	0.0507	0.0507	0.0632	0.0503	0.0499
maximum	0.0751	0.0521	0.0701	0.0519	0.0514
Míry variability					
směrodatná odchylka	0.0091	0.0021	0.0045	0.0013	0.00078
variační koeficient (%)	16.8	4.3	7.6	2.6	1.6
Míry šikmosti a špičatosti					
šikmost	0.0437	-1.2445	0.2573	-0.3238	0.3029
špičatost	3.9179	3.5649	2.2969	3.1578	2.8933

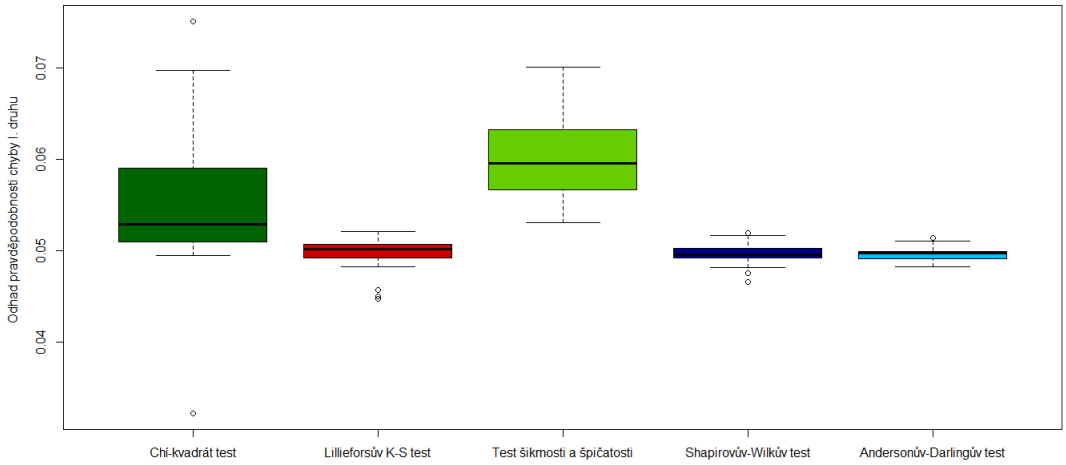
Tabulka 2: Výběrové charakteristiky pro odhad pravděpodobnosti chyby I. druhu

Tabulka 2 obsahující výběrové charakteristiky pro jednotlivé testy normality. Můžeme z ní vyčíst podobně jako z grafu Obrázek 6.1 několik vlastností, které testy mají.

Kde jednotlivé zkratky znamenají: CHI - Chí-kvadrát test dobré shody, KS - Kolmogorovův-Smirnovův test (Lillieforsova varianta), SS - test založený na výběrové šikomosti a špičatosti, SW - Shapirův-Wilkův test normality a AD - Andersonův-Darlingův test.

Podle velikostí pravděpodobností minima a maxima, které jsou u chí-kvadrát testu dobré shody a testu šikmosti a špičatosti usuzujeme, že se jedná o testy nejvíce chybující. Všechny mediány a průměry se pohybují kolem hladiny významnosti. Jediný průměr, který je vyšší než ostatní průměry testů, je právě průměr u testu šikmosti a špičatosti, což svědčí o tom, že tento test je nejhorším testem normality. Naopak nejméně chybujícím je Andersonův-Darlingův test a Shapirův-Wilkův test.

V tabulce také můžeme vidět, že šikmost a špičatost vyšla poměrně rozumně. Špičatosti se pohybují vesměs kolem 3, tzn. že žádný z testovaných výběrů není nějak významně špičatější nebo plošší.



Obrázek 6.2: Vícenásobný krabicový graf pro odhady pravděpodobností chyby I. druhu

Nakonec se můžeme pokusit posoudit, který z testů vykazuje nejmenší četnost chyby I. druhu. Chtěli jsme provést ANOVu, jelikož se ale ukázalo, že nepocházejí všechny data z normálního rozdělení, ANOVu udělat nemůžeme. Na naměřená data proto použijeme Kruskal-Wallisův test, alternativu jednofaktorového testu ANOVA pro výběry, které nepocházejí z normálního rozdělení, případně nesplňují předpoklad homoskedasticity.

Výsledná p-value poukazovala na fakt, že mezi výsledky jednotlivých testů pro ověřování normality jsou statisticky významné rozdíly. K tomu, abychom byli schopni určit, kde se ony rozdíly nacházejí, jsme mohli použít například Wilcoxonův test. Jelikož jsme jím testovali data po párech, prováděli jsme jej pro deset dvojic, s ohledem na to jsme aplikovali Bonferoniho korekci hladiny významnosti, na níž jsme Wilcoxonův test prováděli. Za hladinu významnosti při provádění Wilcoxonova testu tedy bereme $\alpha = \frac{0{,}05}{10} = 0,005$.

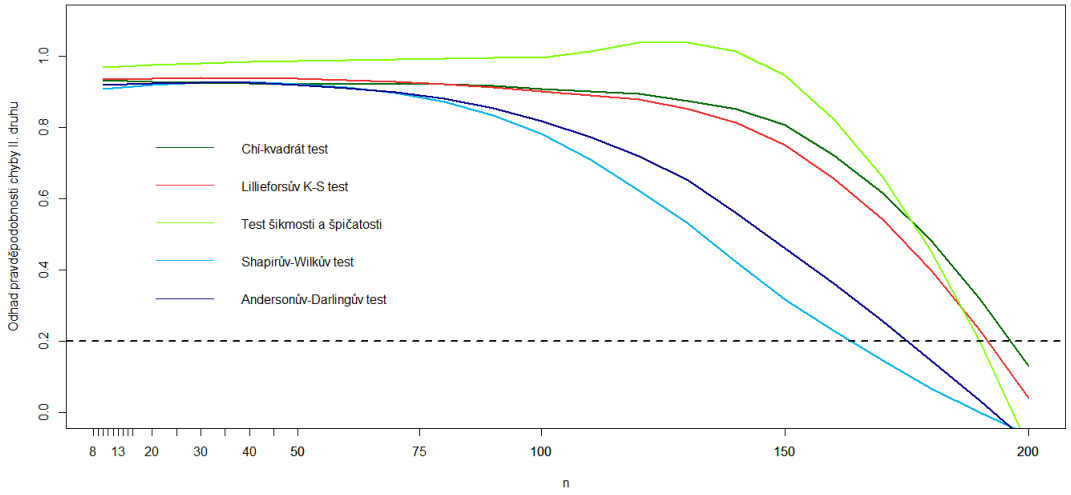
Srovnáním jednotlivých výsledků Wilcoxonova testu pro jednotlivé páry jsme pak došli k závěru, že podle chyby I. druhu je nejméně spolehlivým test šikmosti a špičatosti, dále pak Chí-kvadrát test (test dobré shody), zbylé 3 měly poměrně vyrovnané výsledky, nicméně přece jen se dalo usoudit, že nejlepší výsledky vykazoval Shapiro-Wilkův test normality. Jak jsme se zmínili již dříve, Shapirův-Wilkův test je předepisován i normou ČSN 01 0225 [7].

6.2 Srovnání podle chyby II. druhu

Jak už jsme si jednou řekli, chyba druhého druhu nastane, pokud nezamítneme normalitu u dat, která nejsou výběrem z normálního rozdělení. Srovnání testů podle této chyby

provádíme tedy na výběrech z rovnoměrného, exponenciálního, logaritmicko-normálního a ze Studentova rozdělení a zjišťujeme, kolikrát jsme nezamítli nulovou hypotézu o normalitě dat. Náhodně vygenerovaných dat máme u těchto typů rozdělení stejně, jako při zkoumání chyby I. druhu a to je 1 000 000.

6.2.1 Spojité rovnoměrné rozdělení

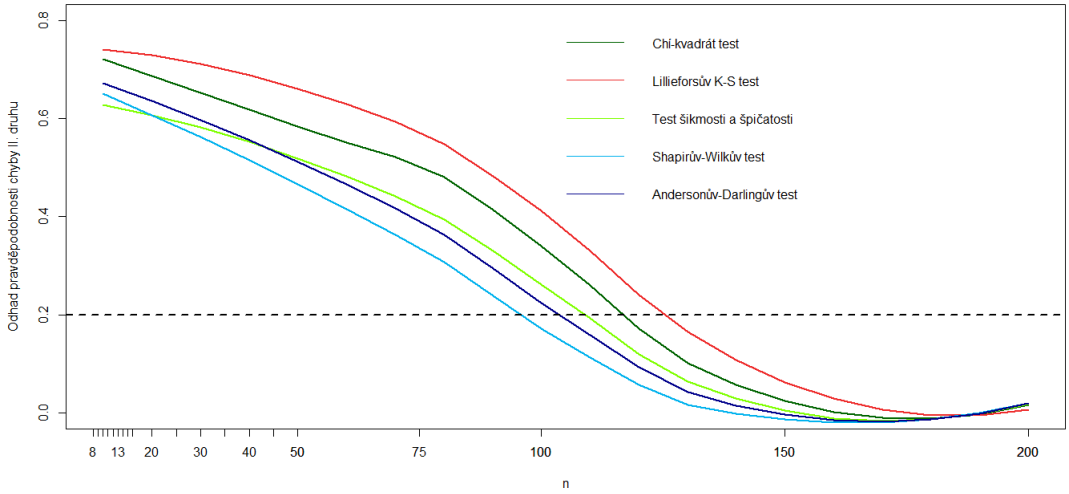


Obrázek 6.3: Odhad pravděpodobnosti chyby II. druhu při rozhodování hypotézy o normalitě dat výběru z rovnoměrného rozdělení

Obrázek 6.3 nám znázorňuje graf, kde máme odhady pravděpodobností chyb II. druhu vzhledem k ověřování nulové hypotézy o normalitě výběru z rovnoměrného spojitého rozdělení na hladině významnosti 0.05.

U malých velikostí výběrů vykazují všechny testy špatné výsledky a velké pravděpodobnosti chyby. Test založený na šikmosti a špičatosti se poměrně dlouho drží na úrovni pravděpodobnosti chyby II. druhu nad 95 %, potom při velikosti výběru 150 prvků začně chyba prudčeji klesat dolů. Nejlépe se však u větších výběrů prokázal Shapirovův-Wilkův test, stejně tak působí test Andersonův-Darlingův. Test Chí-kvadrát a Lillieforsova varianta K-S testu v případě dat z rovnoměrného rozdělení pravděpodobnosti jsou si velice podobné svými odhady pravděpodobností chyby druhého druhu.

6.2.2 Exponenciální rozdělení

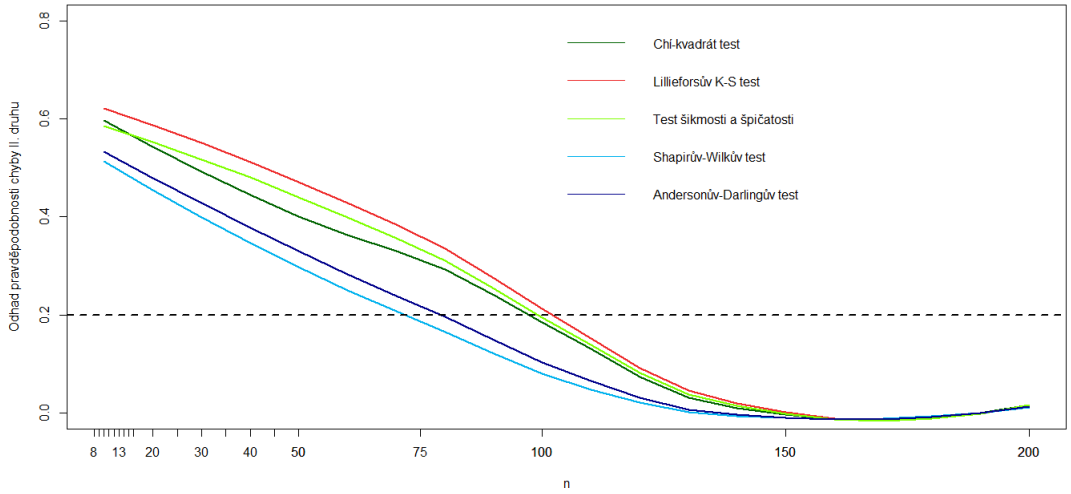


Obrázek 6.4: Odhad pravděpodobnosti chyby II. druhu při rozhodování hypotézy o normalitě dat výběru z exponenciálního rozdělení

Graf s odhady pravděpodobností chyb II. druhu vzhledem k ověřování nulové hypotézy o normalitě výběru z exponenciálního rozdělení pravděpodobnosti na hladině významnosti 0.05 znázorňuje Obrázek 6.4.

Nejvíce chybujícím testem se zdá být Lillieforsův Kolmogorovův-Smirnovův test, u kterého se pravděpodobnosti chyby pohybují nejvíš. Andersonův-Darlingův test s Shapirovým-Wilkovým testem jsou si nejvíce podobní. Ostatní testy mají nulovou chybu až po nich.

6.2.3 Logaritmicko-normální rozdělení

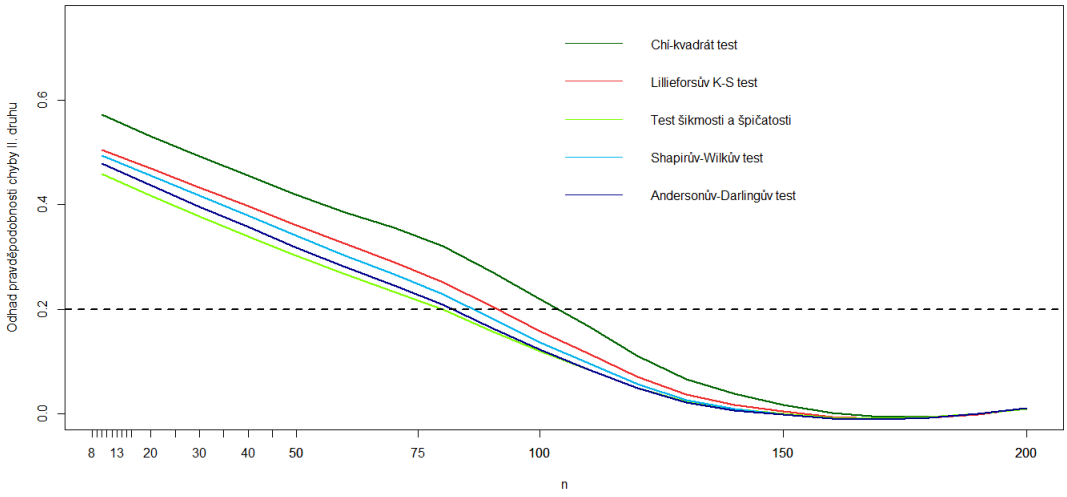


Obrázek 6.5: Odhad pravděpodobnosti chyby II. druhu při rozhodování hypotézy o normalitě dat výběru z logaritmicko-normálního rozdělení

Obrázek 6.5, na kterém můžeme vidět graf odhadů pravděpodobnosti chyby druhého druhu vzhledem k ověřování nulové hypotézy o normalitě výběru z logaritmicko-normálního rozdělení na hladině významnosti 0.05.

Graf vypadá velice podobně jako graf pro výběry z exponenciálního rozdělení, Obrázek **??**, Lillieforsův Kolmogorovův-Smirnovův test jo podle grafu horší než ostatní testy. Opět se zde jeví jako nejlepší test Shapirův-Wilkův, Andersonův-Darlingův test normality se výsledky velice podobá Shapirovu-Wilkovu testu. Test založený na šikmosti a špičatosti společně s testem dobré shody jsou svými výsledky až za Andersonovým-Darlingovým testem.

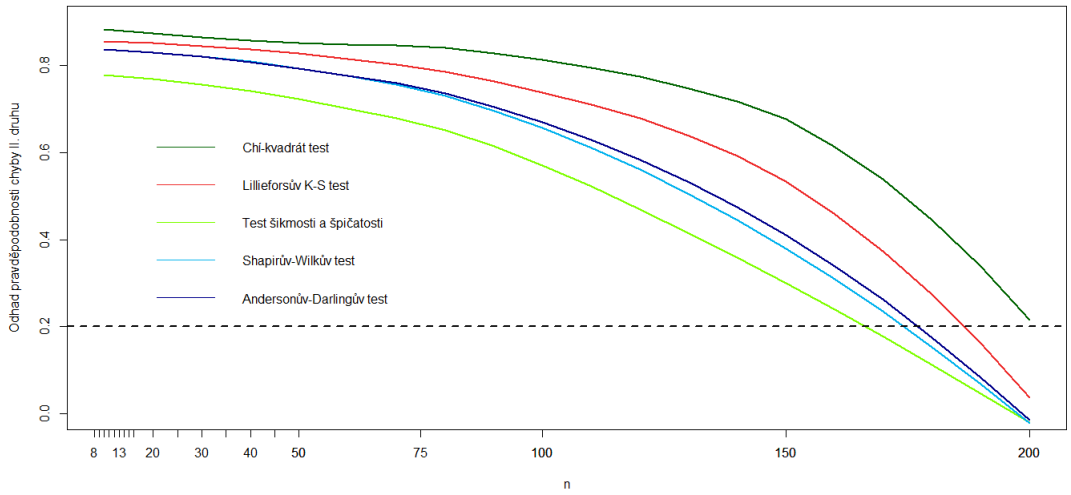
6.2.4 Studentovo rozdělení



Obrázek 6.6: Odhad pravděpodobnosti chyby II. druhu při rozhodování hypotézy o normalitě dat výběru ze Studentova rozdělení s jedním stupněm volnosti

Na obrázku Obrázek 6.6 je znázorněn graf odhadů pravděpodobnosti chyby druhého druhu vzhledem k ověřování nulové hypotézy o normalitě výběru ze Studentova rozdělení s jedním stupněm volnosti na hladině významnosti 0.05.

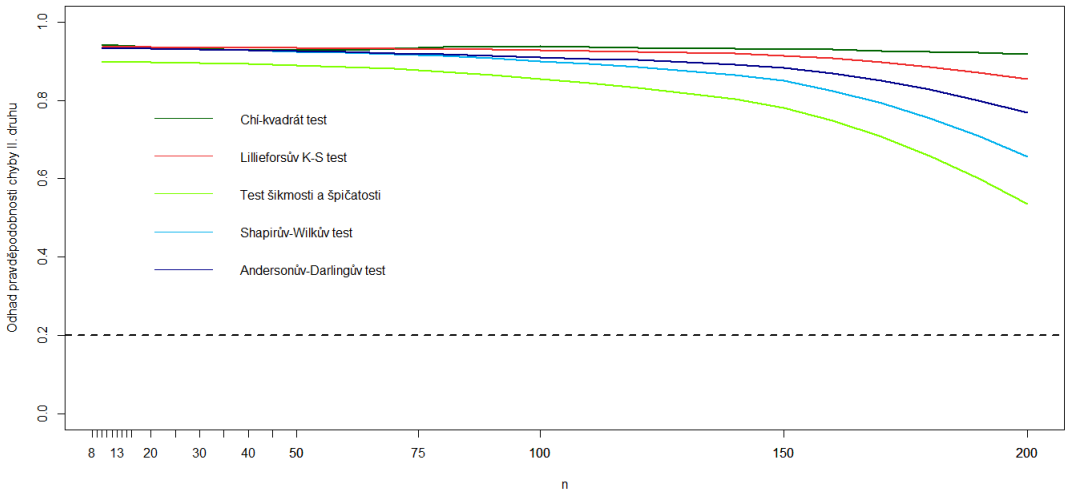
U Studentova rozdělení s jedním stupněm volnosti se nám pozměnilo pořadí testů normality a jako nejlepší se tváří test založený na šikmosti a špičatosti, který má odhad pravděpodobnosti chyby druhého druhu menší než 20 % už při velikosti 75 prvků, těsně za ním je test Andersonův-Darlingův, velice podobně je na tom Shapirovův-Wilkův test. Chí-kvadrát test dobré shody dopadl ze všech testů nejhůř.



Obrázek 6.7: Odhad pravděpodobnosti chyby II. druhu při rozhodování hypotézy o normalitě dat výběru ze Studentova rozdělení se třemi stupni volnosti

Graf Obrázek 6.7 zobrazuje odhady pravděpodobností chyb II. stupně vzhledem k ověřování nulové hypotézy o normalitě výběru ze Studentova rozdělení pravděpodobnosti se třemi stupni volnosti na hladině významnosti $\alpha = 0.05$.

Zde dopadl nejhůř test dobré shody Chí-kvadrát, po něm následuje Lillieforsova varianta Kolmogorova-Smirnova testu normality. Při malých velikostech výběrů mají totožné výsledky test Shapirův-Wilkův a Andresonův-Darlingův test, jenže u výběru velikosti 100 prvů se stává Shapirův-Wilkův test lepším. Nejlepším testem u Studentova rozdělení se 3. stupni volnosti je test založený na šikmosti a špičatosti.



Obrázek 6.8: Odhad pravděpodobnosti chyby II. druhu při rozhodování hypotézy o normalitě dat výběru ze Studentova rozdělení s deseti stupni volnosti

Posledním grafem, kterým zobrazujeme odhady pravděpodobností chyb II. druhu vzhledem k ověřování nulové hypotézy o normalitě výběru ze Studentova rozdělení s deseti stupni volnosti na hladině významnosti 0.05, je graf na obrázku Obrázek 6.8.

Pořadí testů zde máme stejné jako u předchozích dvou grafů. Můžeme si všimnout, že čím více stupňů volnosti Studentovo rozdělení má, tím obtížnější je odlišit ho od normálního rozdělenení. Při výběru dat ze Studentova rozdělení s deseti stupni volnosti, nedokážeme podle grafu říct při jaké velikosti prvků bude odhad pravděpodobnosti chyby II. druhu menší než 20 %.

Chyba druhého druhu se zvyšuje s rostoucím stupněm volnosti, protože čím více stupňu volnosti, tím blíž se podobá Studentovo rozdělení přávě normálnímu rozdělení pravděpodobnosti.

7 Transformace dat vedoucí k přiblížení k normalitě

Pokud se při analýze dat zjistí, že rozdělení datového výběru se příliš liší od normálního rozdělení (odlehlé body, asymetrie, nehomogenita), nastává problém, jak data vůbec vyhodnotit. Pro vyhodnocení dat, lze v řadě případů použít vhodnou *transformaci*, která vede ke zesymetričtění rozdělení, stabilizaci rozptylu a někdy dokonce k normalitě. Výchozí představa je taková, že zpracovávaná data jsou nelineární transformací náhodné veličiny x s normáním rozdělením. Hledá se k nim pak inverzní transformace $g(x)$.

Pokud chceme **stabilizovat rozptyl**, vyžaduje to, abychom nalezli transformaci $y = g(x)$, ve které je již rozptyl $\sigma^2(y)$ konstantní. Pokud je ale rozptyl původní proměnné x funcí $\sigma^2(x) = f_1(x)$, můžeme rozptyl $\sigma^2(y)$ určit jako

$$\sigma^2(y) \approx [\frac{\delta g(x)}{\delta x}]^2 \cdot f_1(x) = C, \tag{7.1}$$

kde C je nějakou konstantou. Potom hledaná transformace $g(x)$ je řešením rovnice

$$g(x) \approx C \int \frac{dx}{\sqrt{f_1(x)}}. \tag{7.2}$$

Pokud je závislost $\sigma^2(x) = f_1(x)$ mocninná, bude optimální transformace $g(x)$ také mocninná. Jelikož je střední hodnota pro normální rozdělení na rozptylu nezávislá, bude transformace, která stabilizuje rozptyl také zajišťovat přiblížení k normalitě.

Prostá **mocninná transformace** nebo také symetrizující transformace zajišťuje zesymetričtění rozdělení výběru a je ve tvaru

$$y = g(x) = \begin{cases} x^\lambda, & \text{pro} \quad \lambda > 0 \\ \ln x, & \text{pro} \quad \lambda = 0 \\ -x^{-\lambda}, & \text{pro} \quad \lambda < 0. \end{cases} \tag{7.3}$$

Optimální odhad λ se hledá minimalizací asymetrie. Kromě klasické šikmosti je možné užít i robustní verzi šikmosti.[11]

7.1 Boxova-Coxova transformace

Tato transformace přibližuje rozdělení výběru k normálnímu vzhledem k šikmosti a špičatosti. Transformace je použitelná pouze pro kladná data a je definovaná

$$y = g(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \text{pro} \quad \lambda \neq 0 \\ \ln x, & \text{pro} \quad \lambda = 0. \end{cases} \tag{7.4}$$

Transformace má tyto vlastnosti:

1. Transformace $g(x)$ jsou vzhledem k veličině λ spojité, protože platí

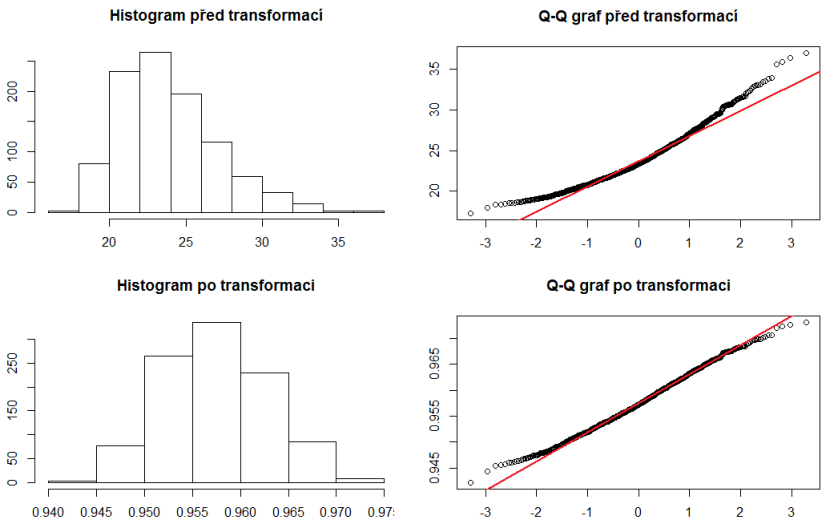
$$\lim_{\lambda \to 0} \frac{x^{\lambda} - 1}{\lambda} = \ln x.$$

2. Všechny transformace procházejí bodem $[x = 0; y = 1]$ a mají v tomto bodě společnou směrnici.

3. Mocninné transformace s exponenty z intervalu $\langle -2; 2 \rangle$ jsou co do křivosti rovnoměrně rozmístěné. [11]

Boxova-Coxova transformace pro výběr dat z kladně zešikmeného rozdělení

Vzali jsme soubor s výběrem dat z kladně zešikmeného rozdělení pravděpodobnosti, který jsme používali již dříve při grafickém ověřování normality a testování statistických testů. Použili jsme Boxovu-Coxovu transformaci, která by měla výběr dat přiblížit více k normálnímu rozdělení.



Obrázek 7.1: Boxova-Coxova transformace dat kladně zešikmeného rozdělení na normální rozdělení

Na obrázku Obrázek 7.1 máme zobrazené histogramy a Q-Q grafy pro data před transformací a po transformaci. Data podle Shapirova-Wilkova testu nesplňují normalitu, ale alespoň se jí přiblížili. P-hodnoty vidíte v Tabulka 4.

Tabulka 3: Výsledky Shapirova-Wilkova testu před a po transformaci

	p-hodnota
Data před transformací	2.73e-16
Data po transformaci	0.002894

Tabulka 4: Výsledky Shapirova-Wilkova testu před a po transformaci

8 Volně šířitelný software pro ověřování normality

Nejdříve jsme chtěli vytvořit volně šířitelný výpočetní applet v softwaru RStudio, usoudili jsme ale nakonec, že bude vhodnější věnovat více času kapitole 6 Srovnání testů podle chyby I. a II. druhu, jelikož jsou již veškeré testy normality v programu RStudio naimplementovány.

V této práci jsem použila již předdefinované funkce jako jsou např. `ad.test()`, `lillie.test()`, `shapiro.test()`,$\cdots$

9 Závěr

Práce je zaměřena zejména na popisnou stránku testů a grafických metod. V rámci přípravy bakalářské práse jsem se blíže seznámila s grafickými metodami pro ověřování romality a se statistickými testy normality.

V konečné části mé bakalářské práce byla provedana simulační studie na vygenerovaných datech z různých typů rozdělení pravděpodobnosti. Testy byly srovnávány na základě odhadů pravděpodobností chyb I. a II. druhu. Na základě těchto simulací nejde jednoznačně určit, který test je nejsilnější, neboť počet datových výběrů by musel být větší. Podle našeho srovnání a normy ČSN 01 0225 [7] si ale troufáme říci, že za nejlepší test lze považovat Shapirův-Wilkův test normality.

Potom jsme si také ukázali, jak je možné některá data z jiného rozdělení transformovat a přiblížit je tak k normálnímu rozdělení pravděpodobnosti.

Veškeré výpočty probíhaly ve statistickém programu Rstudio, který jsme si zvolili, protože je veřejně dostupný.

10 Reference

[1] ANDĚL Jiří, *Statistické metody*, Praha: MatFyzPress, 2007, ISBN 80-7378-001-1.

[2] FROBENSKÁ Marie, KOLÁČEK Jan, *Pravděpodobnost a statistika I* [online], Brno: Masarykova univerzita, 2013, dostupné na: is.muni.cz/elportal/?id=1130308, ISBN 978-80-210-6710-3.

[3] HEBÁK Petr, Diana BÍLKOVÁ a Alžběta SVOBODOVÁ, *Praktikum k výuce matematické statistiky II: testování hypotéz,* Praha: Oeconomica, 2004, ISBN 80-245-0721-8.

[4] HEBÁK Petr, *Testování statistických hypotéz* Praha: Vysoká škola ekonomická, 1995, ISBN 80-7079-294-9.

[5] HENDL Jan, *Přehled statistických metod: analýza a metaanalýza dat,* Praha: Portál, 2009, ISBN 978-80-7367-482-3.

[6] HOLICKÝ Milan, *Aplikace teorie pravděpodobnosti a matematické statistiky,* Praha: České vysoké učení technické, 2015, ISBN 978-80-01-05803-9.

[7] JAROŠ F., ROSA Z., *ČSN 01 0225 - Aplikovaná statistika. Testy shody empirického rozdělení s teoretickým*, Praha: Český normalizační institut, 1980.

[8] LILLIEFORS H. W., *On the Kolmogorov–Smirnov test for normality with mean and variance unknown*, Journal of the American Statistical Association vol. 62, s.399–402, 1967.

[9] LITSCHMANNOVÁ Martina, *Vybrané kapitoly z pravděpodobnosti* [online], Ostrava, 2011, dostupné na: www.am.vsb.cz/litschmannova.

[10] LITSCHMANNOVÁ Martina, *Úvod do statistiky* [online], Ostrava, 2011, dostupné na: www.am.vsb.cz/litschmannova.

[11] MELOUN Milan, *Statistická analýza jednorozměrných dat* [online], Pardubice: Univerzita Pardubice, 2011, dostupné na: www.crr.vutbr.cz/system/files

[12] ZVÁRA Karel,ŠTĚPÁN Josef, *Pravděpodobnosti a matematická statistika*, Praha: Matfyzpress, 2012. ISBN 978-80-7378-218-4.

A Funkce pro test založený na šikmosti a špičatosti

Funkce 1 je naprogramována pro test založený na šikmosti a špičatosti, kde `x` je vstupní vektor s daty, která chceme otestovat. Výstupem fuknce je `test`, což je výsledek testu. Fukncí používáme pro srovnání chyby I. a II. stupně a voláme ji v dalších naprogramovaných funkcích.

```
test.ss = function(x)
{
n = length(x);
D3 = (6*(n-2))/((n+1)*(n+3));
E4 = 3-(6/(n+1));
D4 = (24*n*(n-2)*(n-3))/((n+1)^2*(n+3)*(n+5));
U3 = (skewness(x))^2/D3;
U4 = (kurtosis(x)-E4)^2/D4;
test  = U3+U4;
return(test);
}
```

Výpis 1: Funkce pro testování šikmosti a špičatosti

B Funkce pro srovnání testů na základě chyby I. druhu

Funkci 2 používáme pro výpočet odhadů pravděpodobností chyb I. druhu. Na vstupu zadáváme `pocet.vyberu`, což je číslo s námi požadovaným počtem výběrů. Výstupem je `seznam`, kde jsou uloženy odhady pro všech 5 statistických testů.

```
chyba1 = function(pocet.vyberu)
{
c1 = 0; c2 = 0; c3 = 0; c4 = 0; c5 = 0;
ch = seq(1,20);    ss = seq(1,20);    sw = seq(1,20);    ad = seq(1,20);    ks = seq(1,20);
vektor=c(8,9,10,11,12,13,14,15,16,20,25,30,35,40,45,50,75,100,150,200);
for(i in 1:20)
{
for(j in 1:pocet.vyberu)
{
xx = rnorm(vektor[i],0,1) ;
alfa_ch = pearson.test(x)[2]$p.value #CHI-KVADRAT TEST
alfa_ss = test.ss(xx); #TEST SIKMOST-SPICATOST
alfa_sw = shapiro.test(xx)[2]$p.value; #SHAPIROVUV-WILKUV TEST
alfa_ad = ad.test(xx)[2]$p.value; #ANDERSONUV-DARLINGUV TEST
alfa_ks = ks.test(xx, "pnorm")[2]$p.value; #KOLMOGOROVUV-SMIRNOVOVUV TEST
if(alfa_ch < 0.05){c1=c1+1};
if(alfa_ss > qchisq(0.95,2)){c2=c2+1};
if(alfa_sw < 0.05){c3=c3+1};
if(alfa_ad < 0.05){c4=c4+1};
if(alfa_ks < 0.05){c5=c5+1};
}
ch[i] = c1/pocet.vyberu;
ss[i] = c2/pocet.vyberu;
sw[i] = c3/pocet.vyberu;
ad[i] = c4/pocet.vyberu;
ks[i] = c5/pocet.vyberu;
c1 = 0; c2 = 0; c3 = 0; c4 = 0; c5 = 0;
}
seznam=as.list(seq(1,5));
seznam[[1]]=ch;
seznam[[2]]=ss;
seznam[[3]]=sw;
seznam[[4]]=ad
seznam[[5]]=ks;
return(seznam);
}
```

Výpis 2: Funkce pro testování testů na základě chyby I. druhu

C Funkce pro srovnání testů na základě chyby II. druhu

Funkci 3 používáme pro výpočet odhadů pravděpodobností chyb II. druhu. Na vstupu zadáváme `pocet.vyberu`, což je číslo s námi požadovaným počtem výběrů. Do proměnné `xx` si ukládáme náhodně vygenerovaná data, v tomhle případě, pocházející z rovnoměrného rozdělení pravděpodobnosti. Výstupem je `seznam`, kde jsou uloženy odhady pro všech 5 statistických testů.

```
chyba2rovno = function(pocet.vyberu)
{
c1 = 0; c2 = 0; c3 = 0; c4 = 0; c5 = 0;
ch = seq(1,20);    ss = seq(1,20);    sw = seq(1,20);    ad = seq(1,20);    ks = seq(1,20);
vektor=c(8,9,10,11,12,13,14,15,16,20,25,30,35,40,45,50,75,100,150,200);
for(i in 1:20)
{
for(j in 1:pocet.vyberu)
{
xx = runif(vektor[i],0,1) ;
alfa_ch = pearson.test(xx)[2]$p.value #CHI-KVADRAT TEST
alfa_ss = test.ss(xx); #TEST SIKMOST-SPICATOST
alfa_sw = shapiro.test(xx)[2]$p.value; #SHAPIROVUV-WILKUV TEST
alfa_ad = ad.test(xx)[2]$p.value; #ANDERSONUV-DARLINGUV TEST
alfa_ks = lillie.test(xx)[2]$p.value; #KOLMOGOROVUV-SMIRNOVOVUV TEST
if(alfa_ch > 0.05){c1=c1+1};
if(alfa_ss < qchisq(0.95,2)){c2=c2+1};
if(alfa_sw > 0.05){c3=c3+1};
if(alfa_ad > 0.05){c4=c4+1};
if(alfa_ks > 0.05){c5=c5+1};
}
ch[i] = c1/pocet.vyberu;
ks[i] = c5/pocet.vyberu;
ss[i] = c2/pocet.vyberu;
sw[i] = c3/pocet.vyberu;
ad[i] = c4/pocet.vyberu;
c1 = 0; c2 = 0; c3 = 0; c4 = 0; c5 = 0;
}
seznam=as.list(seq(1,5));
seznam[[1]]=ch;  seznam[[2]]=ks;
seznam[[3]]=ss;  seznam[[4]]=sw;  seznam[[5]]=ad
return(seznam);
}
```

Výpis 3: Funkce pro testování testů na základě chyby II. druhu

Další fuknce pro chyby II. druhu se nachází v elektronické příloze.

D Distribuční funkce normovaného normálního rozdělení pro $z > 0$

z	0	1	2	3	4	5	6	7	8	9
0,0	0,500	0,504	0,508	0,512	0,516	0,520	0,524	0,528	0,532	0,536
0,1	0,540	0,544	0,548	0,552	0,556	0,560	0,564	0,567	0,571	0,575
0,2	0,579	0,583	0,587	0,591	0,595	0,599	0,603	0,606	0,610	0,614
0,3	0,618	0,622	0,626	0,629	0,633	0,637	0,641	0,644	0,648	0,652
0,4	0,655	0,659	0,663	0,666	0,670	0,674	0,677	0,681	0,684	0,688
0,5	0,691	0,695	0,698	0,702	0,705	0,709	0,712	0,716	0,719	0,722
0,6	0,726	0,729	0,732	0,736	0,739	0,742	0,745	0,749	0,752	0,755
0,7	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
0,8	0,788	0,791	0,794	0,797	0,800	0,802	0,805	0,808	0,811	0,813
0,9	0,816	0,819	0,821	0,824	0,826	0,829	0,831	0,834	0,836	0,839
1,0	0,841	0,844	0,846	0,848	0,851	0,853	0,855	0,858	0,860	0,862
1,1	0,864	0,867	0,869	0,871	0,873	0,875	0,877	0,879	0,881	0,883
1,2	0,885	0,887	0,889	0,891	0,893	0,894	0,896	0,898	0,900	0,901
1,3	0,903	0,905	0,907	0,908	0,910	0,911	0,913	0,915	0,916	0,918
1,4	0,919	0,921	0,922	0,924	0,925	0,926	0,928	0,929	0,931	0,932
1,5	0,933	0,934	0,936	0,937	0,938	0,939	0,941	0,942	0,943	0,944
1,6	0,945	0,946	0,947	0,948	0,949	0,951	0,952	0,953	0,954	0,954
1,7	0,955	0,956	0,957	0,958	0,959	0,960	0,961	0,962	0,962	0,963
1,8	0,964	0,965	0,966	0,966	0,967	0,968	0,969	0,969	0,970	0,971
1,9	0,971	0,972	0,973	0,973	0,974	0,974	0,975	0,976	0,976	0,977
2,0	0,977	0,978	0,978	0,979	0,979	0,980	0,980	0,981	0,981	0,982
2,1	0,982	0,983	0,983	0,983	0,984	0,984	0,985	0,985	0,985	0,986
2,2	0,986	0,986	0,987	0,987	0,987	0,988	0,988	0,988	0,989	0,989
2,3	0,989	0,990	0,990	0,990	0,990	0,991	0,991	0,991	0,991	0,992
2,4	0,992	0,992	0,992	0,992	0,993	0,993	0,993	0,993	0,993	0,994
2,5	0,994	0,994	0,994	0,994	0,994	0,995	0,995	0,995	0,995	0,995
2,6	0,995	0,995	0,996	0,996	0,996	0,996	0,996	0,996	0,996	0,996
2,7	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997
2,8	0,997	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998
2,9	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,999	0,999	0,999
3,0	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
3,1	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
3,2	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
3,3	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Převzato z [9].